

# Conservation genomic management of two critically endangered New Zealand birds.



Stephanie Joy Galla

This thesis is presented for the degree of:

Doctor of Philosophy in Biological Sciences

University of Canterbury  
School of Biological Sciences

2019

# Abstract

In order to conserve global biodiversity, a multifaceted approach is needed to address complex conservation issues. One valuable tool in this approach is the use of genetic data to inform management (i.e., conservation genetics). For intensively managed threatened populations, genetic diversity can be managed through a conservation breeding approach, where relatively unrelated individuals are paired together to minimise inbreeding and maximise diversity in an effort to maximise evolutionary potential. For many, the pedigree has been the tool of choice for making pairing recommendations in conservation breeding programmes, as it relies on available ancestry data to estimate kinship —a measure of coancestry or pairwise relatedness— between individuals. However, threatened species recovery programmes often struggle to use this approach when pedigrees are shallow or incomplete. While genetic data (i.e., microsatellites) can measure relatedness for pairing recommendations, emerging evidence indicates this approach lacks precision in genetically depauperate species and more precise measures may be obtained from genomic data (i.e., thousands of single nucleotide polymorphisms, or SNPs). The field of conservation genetics is currently transitioning from using relatively few genetic markers to using thousands of genome-wide SNPs using high throughput sequencing (HTS) technologies. While the emerging field of conservation genomics promises greater precision for population diversity measures, relatively few studies to date have used these technologies, and exemplars are needed to demonstrate how to effectively and efficiently navigate from genetic to genomic technologies for use in conservation genetic management.

This thesis serves as one such exemplar, using two critically endangered birds as Proof-of-Concept: the kakī/black stilt (*Himantopus novaezelandiae*) and kākāriki karaka/orange-fronted parakeet (*Cyanoramphus malherbi*). Both species are endemic to Aotearoa New Zealand and part of their management includes conservation breeding programmes, where individuals are bred in captivity with their offspring translocated to predator-controlled wild habitats. Pairing recommendations for captive kakī and kākāriki karaka have been based loosely on visualised pedigree diagrams, but no studies to date have formally analysed either pedigree. In order to establish the capabilities and limitations of existing tools for genetic/genomic management, in **Chapter 2** I developed multigenerational pedigrees for both species to investigate founder representation, relatedness, and mean kinship. This chapter highlights limitations of pedigrees for species with conservation breeding programmes that are routinely augmented by individuals of unknown ancestry, and underscores the value in incorporating empirical data (i.e., genetics and genomics) into management.

In the form of a *Molecular Ecology* opinion piece lead by me, **Chapter 3** provides an overview of the gap between the availability of genomic tools and their use for conservation (i.e., the ‘conservation genomics gap’) and provides a pathway for people to transition and upskill in bioinformatic capacity. This piece describes how interdisciplinary relationships are enabling advances in both conservation genomics and primary industry research (e.g., agriculture, fisheries, forestry and horticulture), given the shared goals and applied nature of both disciplines. While conservation geneticists can learn about genomic approaches for aligned questions from primary industry, conservation geneticists can lend biodiversity expertise to primary industry for improved primary production output.

In **Chapter 4**, in an invited submission for the *Genes “Conservation Genetics and Genomics”* Special Issue, my co-authors (including co-first author Natalie Forsdick) and I explore the capacity for using readily available closely-related reference genomes for conservation management. In this chapter, we compare diversity estimates (i.e., nucleotide diversity, individual heterozygosity, and relatedness) derived from SNPs discovered using genotyping-by-sequencing and whole genome resequencing reads mapped to conordinal (killdeer, *Charadrius vociferus*), confamilial (pied avocet, *Recurvirostra avosetta*), congeneric (pied stilt, *Himantopus himantopus*) and conspecific reference genomes. Results indicate that diversity and individual heterozygosity estimates calculated from SNPs discovered using closely related reference genomes correlate significantly with estimates calculated from SNPs discovered using a conspecific genome, with congeneric and confamilial references provide higher correlations and more similar measures. While conspecific genomes may be necessary to address other questions in conservation, SNP discovery in birds using high-quality reference genomes of closely related species is a cost-effective approach for estimating diversity measures in threatened species.

In **Chapter 5**, in a manuscript submitted to *Evolutionary Applications*, my co-authors and I compare relatedness measures using pedigree, genetic, and genomic approaches for making pairing decisions in two critically endangered birds from Aotearoa with conservation breeding programmes: kakī and kākāriki karaka. This study uses family groups (i.e., parents, offspring, and siblings) to assess methods of estimating relatedness, as first order relationships between parents & offspring and siblings in these conservation breeding programmes are known. Our findings indicate genetic measures of relatedness are indeed the least precise when assessing known



parent-offspring and sibling relationships, with SNPs providing more precision. Our results also show that pairing recommendations are most similar when using pedigrees and SNPs. Overall, these results indicate that in lieu of robust pedigrees, SNPs are the most effective measure of relatedness, which has exciting implications for poorly pedigreed populations worldwide.

Beyond using putatively independent SNPs for estimating relatedness, many researchers are looking to discover the genomic basis underlying maladaptive traits in small populations (e.g., inbreeding depression). While outside the scope of this thesis, **Chapter 6** discusses new avenues for research given rich genomic and pedigree data sets now available for both kakī and kākāriki karaka. We anticipate that population genomic management simulations that balance selection for genome-wide diversity while penalising individuals for carrying maladaptive traits will allow researchers the ability to assess whether this approach enhances recovery in threatened populations. Overall, these combined chapters provide a toolbox for conservation geneticists who are transitioning to genomic technologies, especially for conservation breeding programmes. While this research project uses kakī and kākāriki karaka as focal species, it sits under the umbrella of a forward-thinking conservation genomics initiative that seeks to maximise the genetic diversity of a wide range of threatened species and enhance recovery efforts for species worldwide.

# Acknowledgements

There are countless people to thank for the work presented in this thesis. First, I would like to thank tangata whenua, in particular Te Rūnanga o Ngāi Tahu, Te Ngāi Tūāhuriri Rūnanga, Te Rūnanga o Arowhenua, Te Rūnanga o Waihao, and Te Rūnanga o Moeraki for their support. I especially thank Sarah Wilson, Yvette Couch-Lewis and Maku Davis-Te Maire for their guidance with generating and curating the genomic resources for two critically endangered taonga species.

I gratefully acknowledge my husband Jacob Patton. When we were back in the States and I mentioned the opportunity of moving New Zealand for a PhD in conservation genomics, he said “go for it” without a second thought. Every step of the way he has been supportive, even when I have been the most uncertain, and for that I am forever thankful.

I extend a heartfelt thanks to my Senior Supervisor and mentor, Associate Professor Tammy Steeves. During this process, I always felt that we approached this research project as equals. In turn, this approach provided me with ownership of the work presented here, and the confidence I needed to thrive in an academic space. Thank you for your encouragement, guidance, open door, fierce conviction, and kind heart.

I would like to thank my Associate Supervisor and mentor, Dr. Anna Santure. I greatly appreciate your thoughtful edits along the way, gentle encouragement, and expertise regarding all-things relatedness. Your positive energy, calming presence, and resourcefulness have been a great comfort to me. I also thank Roger Moraga for his enormous bioinformatic efforts, expertise, support, and teachings. I appreciate your kindness and assistance, even with the most basic of questions.

I am grateful to the conservation practitioners that we had the privilege of working with, including members of the Kakī Recovery Group (especially Liz Brown, Simone Clealand,

Cody Thyne, and Richard Maloney) and the Orange-fronted Kākāriki Recovery Group (especially Andrew Grant, Lyndon Slater, and Anne Richardson). Thank you for your help with tissue sampling, my endless questions on pedigrees, and for your inspiring efforts to recover taonga species.

I thank the MBIE End User Working Group, who provided robust feedback on the research provided here, including Lucy Bridgeman, Liz Brown, Greg Byrnes, Rod Hitchmough, Carolyn Hogg, Dave Houston, Troy Makan, Axel Moehrenschlager, Kevin Parker, and Phil Seddon.

I gratefully acknowledge two students whom I had the pleasure of mentoring. Many thanks to Maddie Crawl-Dowgray, our summer student, for her research investigating the number of markers needed for providing confident estimates of relatedness in birds. I thank our honours student, Ashley Overbeek, who researched the accuracy of the kakī wild pedigree using microsatellite markers.

I thank those who provided financial support for this research. I am grateful to the University of Canterbury for the UC Doctoral Scholarship. I thank the Ministry of Business, Innovation and Employment for the Endeavour Fund Smart Ideas Grant, which funded the whole genome sequencing and resequencing research provided here, meetings and conferences with researchers in Aotearoa and overseas, an extension scholarship for my continued work, and also open access publication costs. I thank the Brian Mason Scientific Trust and Mohua Charitable Trust for funding the microsatellite and genotyping-by-sequencing research provided here.

Finally, I would like to thank my whānau at home and in Aotearoa, especially my mother Carol Galla, my father Steve Galla, my sister Andrea Chow, and my dear friends, especially Levi Collier-Robinson, Ilina Cubrinovska, Natalie Forsdick, Seoljong Kim, Jasper

Obico, Aisling Rayne, Jennifer Schori, Rachel Van Heughton, and Jana Wold. Thank you for your steadfast assurance, all the way through.

# Table of Contents

<b>Abstract</b> .....	<b>i</b>
<b>Acknowledgements</b> .....	<b>v</b>
<b>List of Figures</b> .....	<b>xi</b>
<b>List of Tables</b> .....	<b>xiii</b>
<b>Chapter 1: Introduction</b> .....	<b>1</b>
1.1: Thesis Layout .....	1
1.2: Conservation Genetics and Genomics .....	2
1.3: The Conservation Genetics and Genomics Gap .....	4
1.4: Genetic Management in Conservation Breeding Programmes .....	5
1.4.1: Pedigrees .....	7
1.4.2: Genetics .....	9
1.4.3: Genomics .....	10
1.5: Life History and Conservation Efforts for Study Species .....	12
1.5.1: Kakī, or Black Stilt .....	12
1.5.2: Kākāriki Karaka, or Orange-fronted Parakeet .....	16
1.6: Aims and Scope .....	20
1.7: References .....	22
<b>Chapter 2: Pedigree Development and Investigation for Kākāriki Karaka and Kakī Conservation</b> .....	<b>35</b>
2.1: Preamble .....	35
2.1.1: Contribution Statement .....	35
2.2: Department of Conservation Pedigree Reports for Kākāriki Karaka and Kakī .....	36
2.2.1: Abstract .....	36
2.2.2: General Introduction .....	37
2.2.3: Kākāriki Karaka Pedigree Report, July 2018 .....	39
2.2.3.1: Kākāriki Karaka Conservation Breeding History .....	40
2.2.3.2: Kākāriki Karaka Pedigree Development .....	42
2.2.3.2: Kākāriki Karaka Pedigree and Founder Representation .....	44
2.2.3.4: Kākāriki Karaka Pairing Recommendations .....	46
2.2.3.5: Kākāriki Karaka Pedigree Discussion .....	47
2.3.4: Kakī Pedigree Report, August 2019 .....	50
2.3.4.1: Kakī Conservation Breeding History .....	50
2.3.4.2: Kakī Pedigree Construction .....	54
2.3.4.3: Kakī Pedigree Results .....	55
2.3.4.4: Kakī Pedigree Discussion .....	60
2.3.5: Concluding Remarks .....	62
2.3.6: References .....	63
<b>Chapter 3: How to Bridge the Conservation Genomics Gap</b> .....	<b>69</b>
3.1: Preamble .....	69
3.1.1: Contribution Statement .....	70
3.2: Building strong relationships between conservation genetics and primary industry leads to mutually-beneficial genomic advances .....	71
3.2.1: Abstract .....	71
3.2.2: Introduction .....	72

3.2.3: Conservation Genetics and Genomics .....	73
3.2.4: Strong Relationships Lead to Mutually Beneficial Genomic Advances .....	82
3.2.5: Moving Forward.....	91
3.2.6: Acknowledgements.....	95
3.2.7: References .....	95
3.2.8: Data accessibility.....	111
<b>Chapter 4: Adding Tools to the Conservation Genomics Toolbox.....</b>	<b>112</b>
4.1: Preamble.....	112
4.1.1: Contribution Statement .....	112
4.2: <i>Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management.....</i>	<i>113</i>
4.2.1: Abstract.....	114
4.2.2: Introduction .....	114
4.2.3: Materials and Methods.....	119
4.2.4: Results.....	129
4.2.5: Discussion .....	136
4.2.6: Conclusions .....	141
4.2.7: Data Availability .....	141
4.2.8: Supplemental Materials.....	142
4.2.9: Funding .....	142
4.2.10: Acknowledgements.....	142
4.2.11: References .....	143
<b>Chapter 5: Comparing available tools for estimating relatedness and making pairing recommendations. ..</b>	<b>151</b>
5.1: Preamble.....	151
5.1.1: Contribution statement .....	152
5.2: <i>Comparison of pedigree, genetic, and genomic estimates of relatedness for informing pairing decisions in two critically-endangered birds: Implications for conservation breeding programmes worldwide.....</i>	<i>153</i>
5.2.1: Introduction .....	154
5.2.2: Materials and methods.....	161
5.2.3: Results.....	171
5.2.4: Discussion .....	180
5.2.5: Acknowledgements.....	189
5.2.6: References .....	189
5.2.8: Data accessibility.....	198
<b>Chapter 6: General Discussion and Future Directions .....</b>	<b>199</b>
6.1: Brief Synopsis.....	199
6.2: The Conservation Genomics Gap, in Retrospect .....	200
6.3: Pairing Recommendations.....	201
6.4: Pedigrees in the Genomics Era.....	206
6.5: Conservation Genomics, Moving Forward .....	208
6.6: Concluding Remarks .....	212
6.7: References .....	212
<b>Appendix A: List of Conservation Genomics Reviews.....</b>	<b>217</b>
<b>Appendix B: Chapter 3, Published Manuscript.....</b>	<b>221</b>
<b>Appendix C: Chapter 3, Supplemental Materials .....</b>	<b>236</b>

C1: Literature Search .....	236
C1.1: Conservation Genomics .....	236
C1.2: Primary Industry Genomics.....	236
<b>Appendix D: Chapter 4, Published Manuscript .....</b>	<b>238</b>
<b>Appendix E: Chapter 4, Supplemental Materials .....</b>	<b>257</b>
E1: Samples.....	257
E2: Sequencing Costs .....	259
E3: Lithium Chloride DNA Extractions Protocol.....	259
<b>Appendix F: Chapter 4, SNP Discovery Pipelines and R-code .....</b>	<b>261</b>
F.1: GBS SNP discovery pipelines.....	261
F.1.1: Axe .....	261
F.1.2: FastQC v. 0.11.4 .....	261
F.1.7: VCFTools & BCFTools v. 1.9 .....	312
F.2: Resequencing SNP-Discovery Pipelines .....	313
F.2.1: FastQC v. 0.11.4 .....	313
F.2.2: Trimmomatic.....	314
F.2.3: Bowtie2 .....	321
F.2.4: Samtools 1.9 Sort .....	342
F.2.5: BCFTools v. 1.9 Mpileup.....	355
F.2.6 BCFTools & VCFTools v. 1.9 .....	355
F.3 R-Scripts .....	357
F.3.1: KGD - General Script for GBS and Resequencing .....	357
F.3.2: Relatedness Scale Transformation & Mantel Tests.....	357
F.3.3: Descriptive Statistic Comparisons .....	358
F.3.4: Correlations.....	361
F.3.5: Heterozygosity Linear Mixed Effects Model .....	364
<b>Appendix G: Chapter 5, Supplemental Materials .....</b>	<b>366</b>
G.1: COANCESTRY Microsatellite Simulations.....	366
G.2: Compute Specifications for Genome Assembly .....	367
G.3: SNP Filtering .....	367
G.4: SNP-based Relatedness Estimates.....	370
Table G6: Pearson's correlation coefficient between different SNP-based estimates of relatedness in kakī.....	372
G.5: MSI and MK Rank Correlations.....	372
G.5: References:.....	374
<b>Appendix H: Chapter 5, SNP Discovery Pipeline and R-code .....</b>	<b>376</b>
H.1: SNP Discovery Pipeline for Kākāriki Karaka.....	376
H.1.2: FastQC .....	376
H.1.3: TrimGalore.....	377
H.1.4: Bowtie2 .....	382
H.1.5: Samtools v. 1.9 Sort.....	384
H.1.6: Mpileup .....	385
H.2.1: BCFTools v. 1.9 View & Prune.....	386
H.2: SNP Filtering using BCFTools v. 1.9 for Kakī .....	388
H.3: R-Scripts:.....	388

H.3.1: General Script for KGD and Relatedness Scale Transformation .....	388
H.3.2: Boxplots for Parent-Offspring and Sibling Relatedness .....	389
H.3.3: Relatedness Correlations.....	390
H.3.4: MSI Correlations .....	396

## List of Figures

### Chapter 1:

Figure 1. 1: Adult kakī in the conservation breeding facility in Twizel, Aotearoa. Photo courtesy of the author. ....	13
Figure 1. 2: Distribution of kakī from the mid 19th century to late 20th century, adapted from Pierce (Figure 1, 1983). Shared with permission from the journal Notornis. ....	14
Figure 1. 3: Current breeding distribution of kakī. ....	15
Figure 1. 4: Kākāriki karaka adult male at the Isaac Conservation and Wildlife Trust in Christchurch, Aotearoa. Photo used with permission, courtesy of Leonie Hayder. ....	17
Figure 1. 5: Breeding distribution of kākāriki karaka. Distributions are denoted with green circles. An enlarged map shows the most recent populations in Canterbury, Aotearoa. ....	19

### Chapter 2:

Figure 2. 1: Map detailing kākāriki karaka breeding populations: 1 — Hurunui, 2 — Poulter Valley, 3 — Hawdon Valley, 4 — Oruawairua/Blumine Island; release sites: 1 — Hurunui, 4 — Oruawairua/Blumine Island, 5 — Te Pākeka/Maud Island, 6 — Tuhua/Mayor Island, 7 — Te Kāhahu-o-Tamatea/Chalky Island; and current captive breeding facilities: 8 — Auckland Zoo, 9 — Pūkaha Wildlife Centre, 10 — Orana Wildlife Park, 11 — Isaac Conservation and Wildlife Trust). Tuhua, Te Pākeka, and Te Kāhahu-o-Tamatea do not currently sustain breeding populations. Pūkaha Wildlife Centre currently holds individuals available for breeding, but at the time of writing of this thesis, no active breeding pairs.....	41
Figure 2. 2: Visualised diagram of the kākāriki karaka captive pedigree. Black lines denote captive pairs that have produced current captive offspring available for pairing. Coloured boxes behind individuals denote captive pairs that produced offspring released on Oruawairua from 2011-12. Faded birds denote dead individuals. ....	45
Figure 2. 3: Map detailing kakī breeding range, release sites (1 — Glentanner, 2 — Mt. Gerald), and captive breeding/rearing facilities (3 — DOC Kakī Captive Management Centre, 4 — Isaac Conservation and Wildlife Trust) in Aotearoa. ....	51
Figure 2. 4: Pedigree diagrams for individual kakī, maintained by the Department of Conservation.....	52
Figure 2. 5: Number of founders that contributed to descendants across the entire pedigree, with the number of descendants binned into groups. ....	58
Figure 2. 6: Founder representation across all breeding individuals over the past three breeding seasons. Founders with an asterix (*) by their ID have been captured in the captive population over the past three seasons. ....	59

### Chapter 3:

Figure 3. 1: Number of publications using high-throughput sequencing technologies to generate genomic data in conservation (blue line) and primary industry (red line) from 2005-2015. Values for this graph were derived from an ISI Web of Science literature search, using inclusive terminology (see Appendix C for details). Curved lines have been smoothed for ease of interpretation. ....	79
Figure 3. 2 : Simplified schematic detailing how relationships between conservation genetics and primary industry are leading to mutually beneficial outcomes. In black arrows, genomic expertise from primary industry advances conservation genetics, which in turn informs conservation biology and conservation management and policy. In white arrows, biodiversity expertise informs primary industry research, which in turn improves primary production. ....	83



## Chapter 4:

Figure 4. 1: Evolutionary relationships between species with reference genomes used in this proof of concept. The evolutionary tree indicates topology between taxa within the order Charadriiformes. Evolutionary tree is not to scale.....	119
Figure 4. 2: Distribution of different diversity estimates using SNPs discovered with GBS reads mapped against different reference genomes. (A) Nucleotide diversity ( $\pi$ ), (B) individual heterozygosity (HS), and (C) pairwise relatedness (R) not including self-relatedness. ....	131
Figure 4. 3: Scatterplots showing individual point estimates of HS (A–C) and pairwise R estimates (D–F) using GBS reads mapped to different reference genomes. Self-relatedness estimates were not used in this analysis. Trend line in black, with 95% confidence intervals surrounding the trend line in gray. Spearman’s correlation coefficient ( $r_s$ ) provided in the lower right corner of each scatterplot. * indicates significance $p < 0.001$ .....	132
Figure 4. 4: Distribution of different diversity estimates using SNPs discovered with resequencing reads mapped against different reference genomes. (A) Nucleotide diversity ( $\pi$ ), (B) individual heterozygosity (HS), and (C) pairwise relatedness (R). Self-relatedness estimates were not used in this analysis. ....	134
Figure 4. 5: Scatterplots showing individual point estimates of HS (A–C) and pairwise R estimates (D–F) using resequencing reads mapped to different reference genomes. Self-relatedness estimates were not used in this analysis. Trend line in black, with 95% confidence intervals surrounding the trend line in gray. Spearman’s correlation coefficient ( $r_s$ ) provided in the lower right corner of each scatterplot. * indicates significance $p < 0.001$ .....	135

## Chapter 5:

Figure 5. 1: Current breeding distributions of wild kakī (A) and kākārīki karaka (B) in Aotearoa. ....	161
Figure 5. 2: Parent-offspring and full sibling relatedness values derived from pedigree- (pale blue), microsatellite- (medium blue), and SNP-based (dark blue) methods in kakī (top graph) and kākārīki karaka (bottom graph). ....	176
Figure 5. 3: Scatterplots showing relationships between pedigree-, microsatellite-, and SNP-based relatedness estimates in known family groups for kakī and kākārīki karaka. A trend line (black) and 95% confidence intervals (grey) are shown in each comparison.. ....	177
Figure 5. 4: Frequency of MSI scores using pedigree - (pale blue), microsatellite- (medium blue), and SNP-based (dark blue) kinship/relatedness values in kakī and kākārīki karaka. ....	180

## Chapter 6:

Figure 6. 1: A wild pair of kakī and their full clutch of fledged offspring, found in Te Manahuna during the 2017-2018 breeding season. Photo used with permission, courtesy of Jemma Welch, Department of Conservation. ....	203
Figure 6. 2: Schematic of suggested simulation study to determine whether pedigree-, SNP-, or SNP-informed pedigree pairings will result in more genome-wide diversity after many generations. ....	206

# List of Tables

## Chapter 2:

Table 2. 1: Pedigree descriptive statistics for all individuals in the kākāriki karaka pedigree, as produced by PMx. ....	44
Table 2. 2: Founder representation across all individuals in the population, captive breeders, captive offspring currently available for pairing (except Poldark), and individuals released to Oruawairua in 2011-2012...44	
Table 2. 3: Individual genetic details of potential breeding individuals, including inbreeding coefficient (F), mean kinship (MK), and pedigree depth. <i>n</i> denotes the number of full siblings that share the same genetic details for this breeding season. It should be noted that MK fluctuates depending on the pool of individuals analysed within PMx. ....	46
Table 2. 4: Pedigree-based estimates of relatedness (R) between all potential breeders. Individuals highlighted in red have first order relatedness ( $R > 0.5$ , pairing not recommended), orange have greater than half-sibling or double first cousin relatedness ( $R = 0.25-0.49$ , pairing not recommended), pale yellow have greater than first cousin relatedness ( $R = 0.125-0.245$ , pairing not recommended), pale green have greater than first cousin once removed relatedness ( $R = 0.0625-0.125$ , can be paired), and grass green are relatively unrelated ( $R < 0.0625$ , ideal pairing). ....	47
Table 2. 5: The number of individuals included in founder analyses ( <i>n</i> ), the number of wild and captive pairs for each season, and list of individuals that were excluded from founder analyses. ....	55
Table 2. 6: Pedigree descriptive statistics for kakī, as produced by PMx, for all known individuals. *Note, the sex ratio is skewed as most individuals in the population are unknown. ....	55
Table 2. 7: All founders represented throughout the pedigree (Unique ID), along with the year sourced, founder representation (rep.), and number of descendants (Descend.). Founders include individuals of wild parentage (Wild), unknown parentage (UNK), multiple parentage (Mult), and individuals of hybrid parentage, with hybrid node provided in parentheses. ....	57
Table 2. 8: Mean kinship for offspring of breeding pairs from the 2018-2019 breeding season. ....	59

## Chapter 3:

Table 3. 1: Common genomic issues facing conservation genetics and parallel examples addressed by scientists in primary industry. ....	85
--	----

## Chapter 4:

Table 4. 1: Genome assembly metrics for the genomes assembled in this study. ....	129
Table 4. 2: Mapping statistics , single nucleotide polymorphisms (SNPs) discovered, SNP descriptive statistics, and average diversity statistics from genotyping-by-sequencing (GBS) reads mapped to different reference genomes. $\pi$ : nucleotide diversity, $H_s$ : individual heterozygosity, $R$ : pairwise relatedness ( $\pm$ S.D. for each measure). ....	130
Table 4. 3: Alignment rates, single nucleotide polymorphisms (SNPs) discovered, SNP descriptive statistics, and average diversity statistics from resequencing reads mapped to different reference genomes. $\pi$ : nucleotide diversity, $H_s$ : individual heterozygosity, $R$ : pairwise relatedness. ( $\pm$ S.D. for each measure). ....	133

## Chapter 5:

Table 5. 1: Family group sampling strategy used in this study. ....	162
Table 5. 2 Descriptive statistics based on pedigree data, as produced by PMx, including number of individuals, sex ratio (males:females), maximum age, gene diversity, number of founders, number of founder genome equivalents, average inbreeding, average mean kinship, average generation time, % ancestry and analytic known, and effective population size. *The sex ratio for all pedigreed individuals for kakī is biased by a large number of individuals with unknown sex. ....	172
Table 5. 3: Descriptive statistics, including number of alleles, observed heterozygosity (HO), and expected heterozygosity (HE) for microsatellite loci used in this study. Loci from kākāriki karaka that were	

monomorphic (OFK12, OFK 19, OFK21, OFK26, OFK31, OFK33, OFK52, OFK56, OFK58, OFK61) are not included. ....	173
Table 5. 4: Descriptive statistics, including number of SNPs pre- and post- filtering, average depth, average missingness, average nucleotide diversity ( $\pi$ ) $\pm$ SD, average proportion of observed heterozygous SNP sites ( $H_O$ ) $\pm$ SD, average proportion of expected heterozygous SNP sites ( $H_E$ ) $\pm$ SD, and average SNP density (number of SNPs per kilobase) $\pm$ SD. ....	175

# Chapter 1: Introduction

## 1.1: Thesis Layout

This thesis contains 6 chapters in total, including an introductory chapter (here), two brief pedigree reports for the Kakī and Kākāriki Karaka Recovery Groups (**Chapter 2**), a published opinion article in the journal *Molecular Ecology* (first author, **Chapter 3**), a published research article in the journal *Genes* (co-first author, **Chapter 4**), a submitted research article in the journal *Evolutionary Applications* (first author, **Chapter 5**), and a conclusion chapter that may also serve as the basis for prospective postdoctoral research (**Chapter 6**). Many of the chapters here are presented as peer-reviewed articles, and as such many contain stand-alone abstracts, introductions, methods, results, discussions, and references.

Because conservation research is inherently interdisciplinary, I intend to use the pronoun *we* instead of *I* when referring to collaborative endeavours. While all the work presented here is led by myself, the use of *we* is intended to embrace and celebrate the collaborative process that has led to these publications. In order to provide context to each of these articles, and also to specify contributions from co-authors, a preamble and contribution statement are provided before every chapter.

**Chapter 1** provides reflections on the role genetics has played in conservation management (section **1.2**). This will include a review of the emerging field of conservation genomics and the transition to using genomic technologies (sections **1.2** and **1.3**). Narrowing the focus, I explore how genetic and genomic technologies can be used for conservation genetic management of critically-endangered species with conservation breeding programmes (section **1.4**). This includes information on current tools that are used

for pairing recommendations (i.e., pedigrees and genetic markers, sections **1.4.1** and **1.4.2**, respectively) and ones that are being investigated (i.e., genomics, section **1.4.3**). This Introduction elaborates on life history traits of two focal organisms in this thesis: the critically-endangered kakī/black stilt (*Himantopus novaezelandiae*; section **1.5.1**) and kākāriki karaka/orange-fronted parakeet (*Cyanoramphus malherbi*, section **1.5.2**). Finally, this Introduction discusses the aims of this thesis (section **1.6**) and provides a list of references used throughout (section **1.7**).

## 1.2: Conservation Genetics and Genomics

Conservation biology is a discipline that addresses loss of biodiversity through the use of interdisciplinary biological research and close collaboration between conservation researchers and practitioners (Soulé 1985; Meffe & Carroll 1997; Haig *et al.* 2016). Given the alarming global loss of biodiversity in the last few centuries (Butchart *et al.* 2010, Barnosky *et al.* 2011), the precedence for the field of conservation biology is momentous and scientists are using tools from multiple disciplines (e.g., ecology, physiology, biogeography, genetics, and social sciences) to answer questions that will inform conservation and restoration decisions (Soulé 1985).

Since its inception more than thirty years ago, subdisciplines within conservation biology have formed which use specific technologies, including the subdiscipline of conservation genetics, which uses genetic data and associated analyses to inform conservation (Avice 2008, Frankham 2010). To date, conservation genetics has used DNA data to identify conservation units (e.g., Serrouya *et al.* 2012), detect population structure (e.g., Schulwitz *et al.* 2014), measure inbreeding (e.g., Brekke *et al.* 2010), relatedness (e.g., Woolaver *et al.* 2013), and diversity at neutral (e.g., Sunny *et al.* 2014) and functional sites

(e.g., Grueber *et al.* 2015) within threatened populations to inform management. After experiencing sudden and drastic population decline (i.e., demographic bottlenecks), threatened populations often experience a reduction in genetic diversity (i.e., genetic bottlenecks; e.g., Johnson & Dunn 2006; Ewing *et al.* 2008; Brekke *et al.* 2010; Bergner *et al.* 2016) and increased inbreeding (i.e., matings between related individuals; Charlesworth & Charlesworth 1987; see also Bouzat 2010; Heber & Briskie 2010), which can contribute to extinction risk (Frankham 2005; Fagan & Holmes 2006). Conservation genetics has allowed for the opportunity to measure and manage the remaining diversity in threatened populations through strategic conservation breeding (Robert 2009) and/or translocation events (Seddon 2010) in an effort to maximise diversity (Frankham 2005) and the ability of populations to adapt to environmental change (i.e., evolutionary potential, de Villemereuil *et al.* 2019).

While the field of conservation genetics began with the use of relatively few markers (e.g., allozymes, restriction fragment length polymorphisms, microsatellites, mitochondrial or nuclear gene sequences; Wan *et al.* 2004), it is now shifting towards the use of many thousands of markers using genomic technologies (Allendorf 2017). Although not formally defined, conservation genomics can be described as the use of genomic technologies to inform conservation management. These technologies may include — but are not limited to — the use of genotyping arrays that produce hundreds of thousands of single-nucleotide polymorphisms, or SNPs (e.g., Minias *et al.* 2019), reduced representation genomics where a fraction of the genome is sequenced through the use of restriction enzymes (e.g., RAD-seq, genotyping by sequencing; Narum *et al.* 2013), or whole genome resequencing where the entire genome is sequenced (Fuentes-Pardo & Ruzzante 2017). The transition towards conservation genomics is accelerated by advances in high-throughput sequencing (HTS)

technologies over the past two decades (Koboldt *et al.* 2013), which continues to reduce the cost associated with producing genomic data (Hayden 2014; Fuentes-Pardo & Ruzzante 2017). In addition to providing more precision to conservation questions regarding population diversity (e.g., Attard *et al.* 2018) and structure (Funk *et al.* 2012), conservation genomics also allows researchers to ask new questions regarding the genomic basis of functional or adaptive traits (e.g., Prince *et al.* 2017) or inbreeding depression (as reviewed in Kardos *et al.* 2016). Indeed, as of 2016 there were 67 reviews, opinion pieces, and book chapters that heralded a conservation genomics approach (see Appendix A for the complete list), which eclipses the number of actual peer-reviewed research articles using conservation genomics that were available at the time (see **Appendix C**; Galla *et al.* 2016; **Chapter 3**).

### 1.3: The Conservation Genetics and Genomics Gap

While conservation genetics and genomics offers great promise, it remains an underutilised resource by the conservation community. For example, a recent survey of 300+ threatened species management plans from Australia, Europe, and the United States of America revealed that only roughly 50% of plans prioritise or include genetic management, with genetics utilised less often in plant than animal management plans (Pierson *et al.* 2016). This paucity in using genetic and genomic data for conservation has been termed the *conservation genetics and/or genomics gap* (Shafer *et al.* 2015; Taylor *et al.* 2017). In the past four years, many have identified contributing factors towards the conservation genetics and genomics gap, and a rich body of literature has developed to create tools to expedite the transition towards genetic and genomic technologies. For example, some have surmised that poor communication and/or a lack of trusted relationships between academic researchers and conservation practitioners (i.e., the research-implementation gap; Knight *et al.* 2008) may hinder the translation of conservation

genetic or genomic research to practice (Shafer *et al.* 2015; Taylor *et al.* 2017). Many have provided helpful suggestions to remedy this hindrance, including the integration of genetic research talks into larger conservation conferences (Taylor & Soanes 2016), the inclusion of practitioners as co-authors on research manuscripts (Britt *et al.* 2018, but see also Hogg *et al.* 2018), and garnering molecular expertise for practitioners through engagement and improved communication (Taylor *et al.* 2017; Hogg *et al.* 2017, 2018; Fabian *et al.* 2019). Regarding the conservation genomics gap, others have identified issues including the high cost of genomic data generation, production of reference genomes, and data storage (Muir *et al.* 2016; Galla & Forsdick *et al.* 2019; **Chapter 4**). The transition between genetics and genomics may also be tempered by the time needed for researchers to upskill towards generating and analysing massive data sets (McCormack *et al.* 2013; Shafer *et al.* 2015; Galla *et al.* 2016; **Chapter 3**).

Although conservation geneticists generally agree that a genetics or genomics gap exists, a literature review has revealed that the number of conservation genomic papers available is on the rise (Galla *et al.* 2016; **Chapter 3**; **Appendix C**). Further, a recent survey of conservation practitioners in Aotearoa New Zealand has revealed that many practitioners understand the value of a genetics and genomics approach to conservation and want to use it in future projects (Taylor *et al.* 2017). While conservation genetics and genomics research is applicable to species management programmes worldwide, it is particularly relevant for Aotearoa, as there are many small, isolated, threatened, and endemic populations that can benefit from strategies to maximise diversity and enhance recovery efforts (Jamieson 2015).

## 1.4: Genetic Management in Conservation Breeding Programmes



Threatened species management programmes often utilise a multi-faceted approach to address factors that cause species decline (e.g., habitat loss, pressure from invasive species, or disease; Wilcove *et al.* 1998) and enhance those that promote recovery (e.g., demographic and genetic management; Jamieson 2015). One tool for demographic and genetic management is conservation breeding, where individuals from threatened populations are removed from wild habitats to prevent extinction and translocated to new environments (e.g., captive facilities or wild sanctuaries) where they can recover in the absence of pressures that caused initial population decline (Fraser 2008, Ballou *et al.* 2010). The goal of this approach is to maintain the diversity of founding individuals and avoid inbreeding by strategically pairing unrelated and underrepresented individuals. In doing so, the population can avoid any negative fitness consequences associated with inbreeding (i.e., inbreeding depression, Charlesworth & Charlesworth 1987) and have sufficient evolutionary potential to adapt to environmental change (Fraser 2008; Ballou *et al.* 2010; Willoughby *et al.* 2015).

A conservation breeding approach has been used by many well-known recovery programmes, including those for the California condor (*Gymnogyps californianus*; Walters *et al.* 2010), black-footed ferret (*Mustela nigripes*; Biggins *et al.* 1999), and golden-lion tamarin (*Leontopithecus rosalia*; Stoinski *et al.* 2003). Worldwide, more than 350 vertebrate species are captive bred for translocation to the wild (Smith *et al.* 2011, 2012), including 20+ such programmes found in Aotearoa (e.g., tuturuatu/shore plover, Davis 1994; kākāriki karaka/parakeet, Ortiz-Catedral *et al.* 2009; whio/blue duck, Glaser *et al.* 2010; kakī/black stilt, Hagen *et al.* 2011; pateke/brown teal, Bowker-Wright *et al.* 2012; kākāpō, Bergner *et al.* 2014). While many species on the brink of extinction have benefitted from conservation

breeding, this approach can be hampered by genetic drift and adaptation to captive environments (Araki *et al.* 2007; Frankham 2008; Willoughby *et al.* 2015; Grueber *et al.* 2017), which may in turn decrease future translocation success (Snyder *et al.* 1996). Therefore, it is important for conservation breeding populations to only remain in captivity for a minimal number of generations without supplementation (Williams & Hoffman 2009) and prioritise the strategic pairing of individuals to maximise diversity (Ballou *et al.* 2010).

#### *1.4.1: Pedigrees*

To maximise diversity, conservation breeding programmes often make pairing decisions to simultaneously avoid inbreeding between pairs and prioritise individuals with low mean kinship, or the average relatedness between an individual to all others in the population, including oneself (Ralls & Ballou 1986; Rudnick & Lacy 2008; Willoughby *et al.* 2015). To achieve this, many conservation breeding programmes are informed by inbreeding and kinship values derived from multigenerational pedigree data (Lacy 2012). Using Mendelian inheritance theory, pedigrees can estimate the probability that a proportion of alleles inherited by an individual are identity-by-descent (IBD) from a parent or common ancestor (Lacy 1995). Pedigrees are often considered the ‘gold standard’ tool for making pairing recommendations in conservation breeding programmes (Jones & Wang 2010; Goudet *et al.* 2018), as they are based on well-understood principles of Mendelian inheritance and supported by an extensive body of literature demonstrating how minimising inbreeding and prioritising individuals with low mean kinship best maximises diversity (Ballou 1983; Ballou & Lacy 1995; Rudnick & Lacy 2008; Lacy 2009; Ballou *et al.* 2010; Ivy & Lacy 2012; Putnam & Ivy 2014; Willoughby *et al.* 2015). Further, software is readily available to manage studbooks (e.g., *SPARKS*, Species360 2017; *PopLink*, Faust *et al.* 2018; *ZIMS*,

Species360 2019) and analyse pedigrees (e.g., *PMx*, Lacy *et al.* 2012), which makes this approach accessible to programmes that collect parentage data during intensive day-to-day management.

While pedigrees are the tool of choice for most conservation breeding programmes (Ivy & Lacy 2012), there are caveats to this approach that hinder their accuracy. For example, pedigrees are based on the assumption that all founders are unrelated. However, founders are likely to be related when they are sourced from threatened wild populations that experienced demographic bottlenecks (e.g., Bergner *et al.* 2014; Hogg *et al.* 2018). While simulation studies have shown that pedigrees are still robust enough to reflect true relatedness and inbreeding values despite this assumption, especially when pedigrees are  $\geq 5$  generations deep (Balloux *et al.* 2004; Pemberton 2004; Rudnick & Lacy 2008), a recent study in Tasmanian devils highlights how violating this assumption led to a significant underestimation of inbreeding (Hogg *et al.* 2018). Many conservation breeding programmes also struggle to use pedigrees when there is missing information from unknown parents, as often seen in species where flock or herd mating is required (Ivy *et al.* 2016). Because pedigrees are often managed by multiple institutions over many years, there is also an opportunity for data entry errors to compound over several generations, leading to inaccurate estimates of relatedness and inbreeding (e.g., Hammerly *et al.* 2016). Even when pedigrees are complete (i.e., containing no unknown, or missing individuals) with no errors, expected relatedness between individuals can differ from realised relatedness, given that pedigrees are based on probabilities as opposed to realised levels of genome sharing (Hill & Weir 2011; Speed & Balding 2015). Because of these caveats, empirical estimates of relatedness based on genetic or genomic information have the potential to better inform management.

### 1.4.2: Genetics

An alternative approach for making pairing recommendations in conservation breeding programmes is to use microsatellite markers (Pemberton 2004; Slate *et al.* 2004; Pemberton 2008). In some instances, these markers are used to reconstruct pedigrees (e.g., Ivy *et al.* 2009) and in other instances they are used to substitute kinship by making empirical estimates of relatedness and/or inbreeding (e.g., Tzika *et al.* 2009). Microsatellite markers have been commonly used for making empirical estimates of relatedness, as they are codominant and highly variable (Weir *et al.* 2006). Furthermore, there are readily available software programmes to estimate pairwise relatedness using microsatellites (e.g., COANCESTRY, Wang 2011) or reconstruct parentage (e.g., CERVUS or COLONY; Marshall *et al.* 1998; Wang 2013). Indeed, there are many programmes that have recently used microsatellites to inform pairing recommendations, repair studbooks, and resolve unknown parentage assignments, including programmes for the lesser kestrel (*Falco naumanni*; Alcaide *et al.* 2010), the critically-endangered Attwater's prairie-chicken (*Tympanuchus cupido attwateri*; Hammerly *et al.* 2013, 2016), the vulnerable Jamaican yellow boa (*Epicrates subflavus*; Tzika *et al.* 2009), and the near-threatened parma wallaby (*Macropus parma*; Ivy *et al.* 2009). In the case of the Attwater's prairie-chicken, shifting from pedigree to genetic-based pairing recommendations has resulted in reduced parental relatedness and a significant increase in chicks living to five weeks of age (Hammerly *et al.* 2016), which demonstrates the great potential of genetic markers for species recovery and conservation.

While many researchers have used microsatellites to inform conservation breeding, more recent simulation studies indicate that genetic-based measures of relatedness based on microsatellites are relatively poor indicators of genome-wide diversity and may provide poor indicators of relatedness and inbreeding, particularly in threatened species where allelic diversity is low (i.e., < 4 alleles per locus in the founding population; Robinson *et al.* 2013; Taylor 2015; Taylor *et al.* 2015). Recent publications suggest that more precise estimates of genome-wide diversity can be obtained from thousands of genome-wide SNPs (Taylor 2015; Taylor *et al.* 2015; Attard *et al.* 2018; Thrasher *et al.* 2018; Benjelloun *et al.* 2019; Lemopoulos *et al.* 2019).

### 1.4.3: Genomics

Given the decreasing cost of high-throughput sequencing (HTS) data production (Hayden 2014), producing thousands of SNPs across multiple samples in a non-model population is now possible. While early genomic studies genotyped small panels of SNPs using SNP arrays for estimating relatedness (e.g., Ross *et al.* 2014), many have now moved towards reduced representation genomic approaches (e.g., genotyping-by-sequencing, or RADseq; Narum *et al.* 2013) or whole genome resequencing (Fuentes-Pardo & Ruzzante 2017; **Chapters 4 and 5**) to discover thousands of SNPs for use in estimating relatedness. While these markers can be discovered without a reference genome (i.e., *de novo* SNP discovery), the use of a reference typically allows for the discovery of higher confidence markers (i.e., higher depth and lower missing data; **Chapter 4**), which may be helpful for providing estimates of relatedness that are not biased by missingness (i.e., missing data; Dodds *et al.* 2015; Attard *et al.* 2018). While there are relatively few conspecific reference genomes available for non-model, threatened species, the number of available

reference genomes is growing exponentially (Ellegren 2014), and at the time of conception of this thesis, it was unknown whether references from closely-related species could be used as a proxy for marker discovery (as per Card *et al.* 2014; see also **Chapter 4**), which would to a faster transition towards genomic approaches.

The number of studies using genome-wide SNPs to estimate relatedness is sparse but growing (see Attard *et al.* 2018). This paucity is likely due in part to the recent availability of affordable HTS technologies or the lack of bioinformatic pipelines and expertise needed for HTS data assembly and management in conservation biology (Shafer *et al.* 2015; **Chapter 3**). Furthermore, there is debate on which approach is best for estimating relatedness (Speed & Balding 2015; Attard *et al.* 2017). Still, genomic technologies offer an exciting opportunity to improve conservation breeding efforts (He *et al.* 2016) and there are examples where SNPs have been used to estimate relatedness and inbreeding, including a studies in zebra finch (Santure *et al.* 2010; Forstmeier *et al.* 2012), field mice and harbor seals (Hoffman *et al.* 2014), the Chat Murciano and Iberian breeds of pig (Herrero-Medrano *et al.* 2012; Silio *et al.* 2015), and Finnish and soay sheep (Li *et al.* 2011; Bérénos *et al.* 2014). Further, a growing number of studies show genome-wide SNPs provide greater accuracy in estimating relatedness and inbreeding over pedigrees (Santure *et al.* 2010; Li *et al.* 2011; Bérénos *et al.* 2014; Kardos *et al.* 2015; Wang 2016) and microsatellites (Hoffmann *et al.* 2014; Hellmann *et al.* 2016; Attard *et al.* 2018; Thrasher *et al.* 2018; Lemopoulos *et al.* 2019; **Chapter 5**).

With that said, exemplars are needed to demonstrate how to efficiently transition towards genomic approaches for estimating relatedness in non-model species. Because relatively few studies have been published comparing approaches for estimating

relatedness and how they affect pairing recommendations (but see Ivy *et al.* 2016; Hogg *et al.* 2018), a Proof-of-Concept is also warranted to compare all approaches available (i.e., pedigree-based, microsatellite-based, and SNP-based). In Aotearoa, two excellent candidates for this study are the kakī/black stilt (*Himantopus novaezelandiae*) and kākāriki karaka (*Cyanoramphus malherbi*), as both have active conservation breeding programmes, multigenerational pedigrees (**Chapter 2**), developed panels of microsatellites (Steeves *et al.* 2008; Andrews *et al.* 2013) and genomic resources available (**Chapter 4** and **Chapter 5**).

## 1.5: Life History and Conservation Efforts for Study Species

### 1.5.1: Kakī, or Black Stilt

Kakī is a critically endangered wading bird endemic to braided river habitats of Aotearoa (Pierce 1996; Figure 1.1). In te reo Māori (i.e., the language of the indigenous people of Aotearoa), kakī translates to neck or throat (<http://maoridictionary.co.nz/>), which likely describes the long neck of the bird. Kakī are considered taonga (i.e., *treasured*) by tangata whenua (i.e., *people of the land*) and were formally recognised as such in Schedule 97 of the Ngāi Tahu Claims Settlement Act of 1998 (<http://www.legislation.govt.nz/>). Although kakī are recognised as taonga, oral or written history regarding Māori cultural connections to kakī is scarce (Nekerangi Paul, *personal comm.*).



Figure 1. 1: Adult kakī in the conservation breeding facility in Twizel, Aotearoa. Photo courtesy of the author.

Once found throughout the North and South Islands of Aotearoa, kakī experienced significant population declines throughout the 20th century due to introduced mammalian predators (e.g., cats, hedgehogs, and ferrets; Reed 1998; Sanders & Maloney 2002) along with habitat loss and degradation (see Figure 1.2, adapted from Pierce 1983). Today, kakī are largely found in Te Manahuna, or the Mackenzie Basin on the South Island (Figure 1.3). In 1981, there were approximately 23 kakī left in the wild and intensive management strategies were initiated by the Department of Conservation to prevent extinction and enhance species recovery (Reed 1998; Steeves *et al.* 2010). These intensive management strategies included predator control efforts (Keedwell *et al.* 2002), management of interspecific hybridisation with poaka (*Himantopus leucocephalus*; Steeves *et al.* 2010; Forsdick *in prep*), invasive plant control (Maloney *et al.* 1999), and a conservation breeding programme (van Heezik *et al.* 2005). The kakī conservation breeding programme includes captive breeding, where individuals are paired and bred in captivity and their offspring are hand-reared or fostered until they are juveniles or sub adults. Another component of the conservation breeding programme is captive rearing, where wild nests are closely monitored and their eggs are harvested, incubated, and hand-reared in captivity until



juvenile/subadult age. Because kakī are socially monogamous, all eggs collected from the wild are assumed to be genetic offspring of putative parents at the nest. However, a recent study has highlighted instances of interspecific egg dumping detected at wild nests (Overbeek *et al.* 2017). Further, a preliminary study using microsatellites and allele mismatch exclusion analysis (Jones & Ardren 2003) over three kakī breeding seasons indicates that a small proportion of kakī are misassigned in the pedigree, due either to egg dumping, extra-pair paternity, or processing errors (Overbeek 2019). Captive bred and reared kakī are translocated to predator-controlled sites in Te Manahuna, where they are supplementally fed and monitored post-release (van Heezik *et al.* 2005).

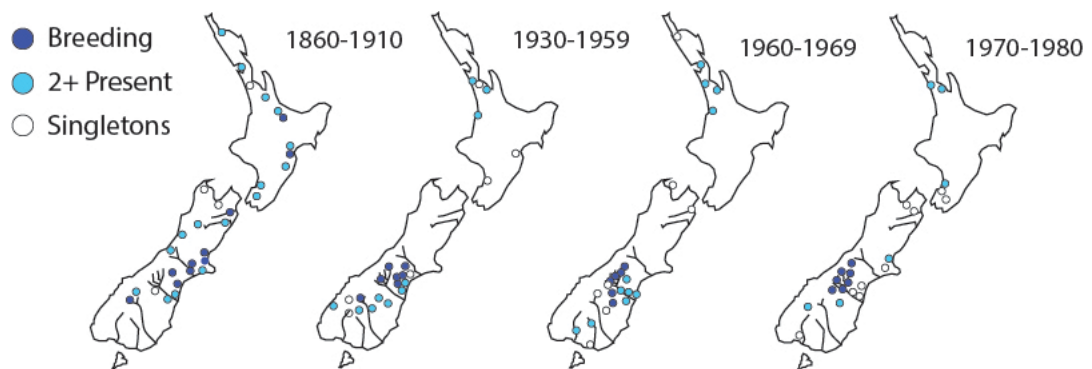


Figure 1. 2: Distribution of kakī from the mid 19th century to late 20th century, adapted from Pierce (Figure 1, 1983). Shared with permission from the journal Notornis.

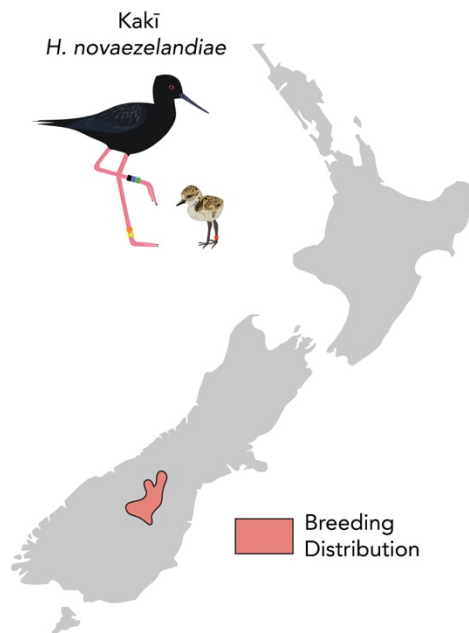


Figure 1. 3: Current breeding distribution of kakī.

There are approximately 129 kakī alive today, and the population is contingent on intensive management efforts (Department of Conservation, *personal comm.*). Recent studies on the conservation breeding population of kakī indicate a negative relationship between breeding pair microsatellite-estimated relatedness and reproductive success (Hagen *et al.* 2011), which demonstrates the importance of effective pairing recommendations to maximise genetic diversity in kakī. A detailed account of life history data and parentage for all managed kakī (called “the alpha list”) has been maintained by the Kakī Recovery Group and the Department of Conservation since the late 1970’s. Further, hand drawn pedigree diagrams for all managed individuals are maintained by the Department of Conservation. Pairing recommendations for kakī usually start with lists of available males and females that appear relatively unrelated and representative of founder diversity based on parentage information from pedigree diagrams. Formal analyses on relatedness using the pedigree have been limited, as it is difficult to perform analyses in its current format (i.e., excel or paper documents) or validate potential data entry errors. Until recently, the pedigree was also presumed to be too shallow or incomplete to provide

accurate estimates of kinship (Rudnick & Lacy 2008), which is another factor contributing to its underutilisation.

To ensure that the potential breeding individuals identified using pedigree diagrams are indeed unrelated, in 2010 the Kakī Recovery Group began using 8 microsatellite markers developed for kakī (Steeves *et al.* 2008) to estimate relatedness between individuals. Individuals with the lowest pairwise relatedness were prioritised for pairing. In kakī, these 8 microsatellite loci have 2-5 alleles per locus (Steeves *et al.* 2008; 2010). Because there is relatively low genetic diversity shown with these markers, they may be a poor indicator of genome-wide diversity in the species (Taylor 2015; Taylor *et al.* 2015). A more appropriate approach for maximising genetic diversity in a genetically impoverished species like kakī may be genomic-based measures of relatedness based on SNPs (Keller *et al.* 2011; Kardos *et al.* 2015; Knief *et al.* 2015; Taylor 2015; Taylor *et al.* 2015; Wang 2016). With the decreasing cost of HTS, the production of genome-wide SNPs are now within reach for kakī (see **Chapters 4 and 5**).

### *1.5.2: Kākāriki Karaka, or Orange-fronted Parakeet*

Kākāriki karaka (Figure 1.4) is also a critically endangered bird endemic to Aotearoa that is considered taonga to tangata whenua, and are recognised along with other parakeets in the genus *Cyanoramphus* in Schedule 97 of the Ngāi Tahu Claims Settlement Act of 1998 (<http://www.legislation.govt.nz/>). In te reo Māori, kākāriki karaka translates to ‘little orange parrot’ (<http://maoridictionary.co.nz/>), which corresponds to the orange frontal band found on the crown. Written history of kākāriki karaka has been limited, as these birds are cryptic, quiet, and inherently difficult to detect (Kearvell *et al.* 2014). Because of their cryptic nature, it is unsurprising that these birds were considered extinct

twice in the 20th century before their 're-discovery' in the Hope Valley of North Canterbury in 1980 (Kearvell *et al.* 2003). The species status of kākārīki karaka has also been shrouded in debate, as they are known to hybridise with sympatric yellow-crowned parakeets (*Cyanoramphus auriceps*); further, these two congeners can be difficult to distinguish in the wild (Kearvell *et al.* 2014). Debate on the taxonomic status has dwindled in the 21st century, as recent field and genetic studies have confirmed that kākārīki karaka are indeed a species-level taxonomic group (Boon *et al.* 2000; Kearvell *et al.* 2003; Rawlence *et al.* 2015) that assortatively mate (Kearvell & Steeves 2015).



Figure 1. 4: Kākārīki karaka adult male at the Isaac Conservation and Wildlife Trust in Christchurch, Aotearoa. Photo used with permission, courtesy of Leonie Hayder.

While kākārīki karaka have been difficult to detect, reports throughout the 19th and 20th centuries have confirmed this species was largely found on the South Island, with a few specimens recorded on the North Island (Kearvell *et al.* 2003). Like many endemic birds to Aotearoa, kākārīki karaka also experienced population declines in the 19th and 20th centuries due to introduced mammalian predators (e.g., rats, brush-tailed possums, stoats) and habitat loss (Kearvell & Legault 2017). Today, an estimated 100-300 kākārīki karaka remain, and their wild breeding populations are restricted to native beech (*Nothofagus spp.*) forests in three valleys of North Canterbury (the Hawdon, the south branch of the

Hurunui, and the Poulter) and Oruawairua/Blumine Island in the Marlborough Sounds, which is a predator free offshore island (Department of Conservation, *personal comm.*; Figure 1.5). Most recent surveys indicate that kākāriki karaka are likely be extinct in the Hawdon Valley of North Canterbury. A few wild individuals have been detected in the Poulter Valley, and many remain in the south branch of the Hurunui, as it is a translocation site for captive bred birds. Recovery efforts for kākāriki karaka include wild nest monitoring, predator control, and a conservation breeding programme. The conservation breeding programme was initiated in 2003 and has continued largely at the Isaac Conservation and Wildlife Trust (Christchurch), Orana Wildlife Park (Christchurch), and the Auckland Zoo (Auckland). Offspring from the conservation breeding programme are all released into the south branch of the Hurunui. To source new founders, a few captive bred offspring have also been released into the Poulter Valley to encourage conspecific pairing with the very few birds in the remnant population. Eggs from these pairings are harvested, with hatchlings being incorporated into the conservation breeding programme.

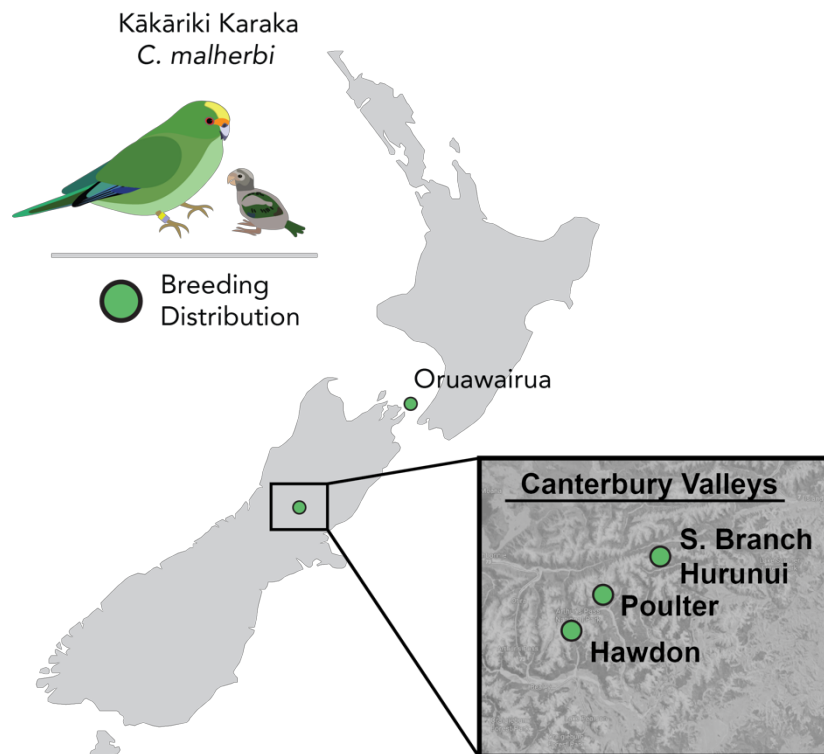


Figure 1. 5: Breeding distribution of kākāriki karaka. Distributions are denoted with green circles. An enlarged map shows the most recent populations in Canterbury, Aotearoa.

A kākāriki karaka studbook has been kept by practitioners at the Isaac Conservation and Wildlife Trust through the studbook keeping programme SPARKS (Species360 2017). While this resource is available, it has not been formally analysed for making pairing recommendations. Instead, pairing recommendations have been informed by ancestry as shown on visualised pedigree diagrams, with an emphasis on maximising founder representation from different valleys in North Canterbury (Anne Richardson, *personal comm.*). While a developed panel of 18 microsatellites exists for kākāriki karaka (Andrews *et al.* 2013), they have not been used to estimate relatedness for captive pairing recommendations. Further, allelic diversity with these markers is low (2-5 alleles per locus; Andrews *et al.* 2013), indicating that genome-wide SNPs may be a more effective approach for making pairing recommendations (**Chapter 5**).

## 1.6: Aims and Scope

The scope of this thesis is to provide resources for how to effectively and efficiently transition towards genomic approaches for the conservation management of threatened species, using two critically-endangered species as Proof-of-Concept: the kakī and kākāriki karaka. Given these two species are taxonomically-distinct, with different life history traits and resources available, they present an ideal opportunity to test the broad-scale applicability of this work.

Neither the kakī nor kākāriki karaka pedigree has ever been formally analysed, and questions regarding founder contribution, pairwise relatedness, and mean kinship have been posed by the Department of Conservation for both species. Therefore, the aim of **Chapter 2** is to develop multigenerational pedigrees in both kakī and kākāriki karaka and investigate questions of interest from the Department of Conservation. In doing this, we revealed instances where the pedigree alone will be insufficient for informing pairing recommendations, especially when considering individuals of unknown ancestry. In these instances, empirical estimates of relatedness, including genomics, will be the most effective for assessing individuals of unknown ancestry.

While genomics offers great promise for greater precision in estimating diversity estimates like relatedness, relatively few conservation researchers (especially when this thesis was initiated in late 2015) have shifted towards genomic technologies, and tools are needed to aid in this transition. The aim of **Chapter 3** is to describe how this transition can be expedited through building mutually-beneficial relationships with another applied biological discipline with aligned genomic research questions: primary industry. This approach builds capacity and capability in the conservation space, as conservation genomic can benefit from genomic expertise and existing pipelines

developed for genomic analyses in primary industry. In return, primary industry research can benefit from biodiversity expertise, which can improve the evolutionary potential of primary production species with low effective population sizes.

To continue building resources for researchers to transition towards genomic technologies, the aim of **Chapter 4** is to explore the capacity for using closely-related reference genomes for SNP discovery and production of neutral diversity estimates, including nucleotide diversity, individual heterozygosity, and relatedness. In this chapter, reduced representation and whole genome resequencing reads for kakī are mapped to conordinal, confamilial, congeneric, and conspecific reference genomes for marker discovery. This chapter demonstrates that diversity estimates calculated from SNPs discovered using closely related reference genomes correlate significantly with estimates calculated from SNPs discovered using a conspecific genome, which indicates that reference genomes from closely related species can be used as a cost-effective approach for estimating diversity measures in threatened bird species.

With tools in hand to transition between genetic and genomic technologies, the aim of **Chapter 5** is to compare relatedness measures using pedigree-, genetic-, and genomic-based approaches for making pairing decisions in critically endangered kakī and kākārīki karaka. Findings from this chapter indicate genetic measures of relatedness are indeed the least precise when assessing known parent-offspring and sibling relationships, with SNPs providing more precision. Our results show that pairing recommendations are most similar when using pedigrees and SNPs, indicating that in lieu of robust pedigrees, SNPs are the most effective measure of relatedness, which has exciting implications for poorly pedigreed conservation breeding programmes worldwide.



Given the rich pedigree and genomic resources developed for kakī and kākārīki karaka, there are more questions beyond neutral diversity that can be asked using a genomics approach. The aim of **Chapter 6** is to explore new potential avenues for research. While this thesis uses kakī and kākārīki karaka as a focal species, it sits under the umbrella of a forward-thinking conservation genomics initiative that seeks to maximise the genetic diversity of a wide range of threatened species with conservation breeding programmes. To ensure this initiative is responsive to diverse end-user needs, the supervisory team on this project have used their extensive networks in the national and international conservation management community, including relevant local rūnanga and iwi trusts, to establish a multi-stakeholder end-user working group. This group has contributed to the wide dissemination and rapid uptake of the initiative by additional New Zealand species recovery programmes (e.g., tuturuatu/shore plover, *Thinornis novaeseelandiae*; tara-iti/fairy tern, *Sternula nereis*; robust grasshopper, *Brachaspis robustus*) and species recovery programmes overseas.

## 1.7: References

- Alcaide M, Negro JJ, Serrano D, Antolín JL, Casado S, Pomarol M (2010) Captive breeding and reintroduction of the lesser kestrel *Flaco naumanni*: A genetic analyses using microsatellites. *Conservation Genetics*, **11**, 331-338.
- Ali OA, O'Rourke SM, Amish SJ, Meek MH, Luikart G, Jeffres C, Miller MR (2016) RAD Capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, **202**, 389-400.
- Allendorf FW (2017) Genetics and the conservation of natural populations: allozymes to genomes. *Molecular Ecology*, **26**, 420-430.
- Andrews BJ, Hale ML, Steeves TE (2013) Characterisation of microsatellite loci in the critically endangered orange-fronted kākārīki (*Cyanoramphus malherbi*) isolated using genomic next generation sequencing. *Molecular Ecology Resources*, **5**, 235-237.
- Araki H, Cooper B, Blouin MS (2007) Genetic effects of captive breeding cause a rapid, cumulative fitness decline in the wild. *Science*, **318**, 100-103.

- Attard CRM, Beheregaray LB, Möller LM (2018) Genotyping-by-sequencing for estimating relatedness in nonmodel organisms: avoiding the trap of precise bias. *Molecular Ecology Resources*, **18**(3), 381-390.
- Awise, J.C. 2008. The history, purview, and future of conservation genetics. In *Conservation biology: evolution in action*. Oxford University Press: 5-15.
- Ballou JD (1983) Calculating inbreeding coefficients from pedigrees. In *Genetics and Conservation*. Cummings Publishing Company: 509-516.
- Ballou JD, Lacy RC (1995) Identifying genetically important individuals for management of genetic variation in pedigreed populations. In *Population management for survival and recovery: analytical methods and strategies in small population conservation*. Columbia University Press: 76-111.
- Ballou JD, Lees C, Faust LJ, Long S, Lynch C, Lackey LB, Foose TJ (2010) Demographic and genetic management of captive populations. *Wild mammals in captivity: principles and techniques for zoo management*. University of Chicago Press: 219-252.
- Balloux F, Amos W, Coulson T (2004) Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, **13**, 3021-3031.
- Barnosky AD, Matzke N, Tomiya S, *et al.* (2011) Has Earth's sixth mass extinction already arrived? *Nature*, **471**, 51-57.
- Benjelloun B, Boyer F, Streeter I, *et al.* (2019) An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity. *Molecular Ecology Resources*, doi:10.1111/1755-0998.13070.
- Béréños C, Ellis PA, Pilkington JG, Pemberton JM (2014) Estimating quantitative genetic parameters in wild populations: A comparison of pedigree and genomic approaches. *Molecular Ecology*, **23**, 3434–3451.
- Bergner LM, Jamieson IG, Robertson BC (2014) Combining genetic data to identify relatedness among founders in a genetically depauperate parrot, the kakapo (*Strigops habroptilus*). *Conservation Genetics*, **15**, 1013-1020.
- Bergner LM, Dussex N, Jamieson IG, Robertson BC (2016) European colonization, not Polynesian arrival impacted population size and genetic diversity in the critically endangered New Zealand kākāpō. *Journal of Heredity*, **107**(7), 593-602.
- Biggins DE, Vargas A, Godbey JL, Anderson SH (1999) Influence of prerelease experience on reintroduced black-footed ferrets (*Mustela nigripes*). *Biological Conservation*, **89**, 121-129.

- Boon WM, Kearvell JC, Daugherty CH, Chambers GK (2000) Molecular systematics of New Zealand Cyanoramphus parakeets: conservation of orange-fronted and Forbes' parakeets. *Bird Conservation International*, **10**, 211-239.
- Bowker-Wright GM, Bell BD, Ritchie P, Williams M (2012) Captive breeding and release diminishes diversity in brown teal *Anas chlorotis*, an endangered New Zealand duck. *Wildfowl*, **62**, 176-189.
- Bouzat J (2010) Conservation genetics of population bottlenecks: the role of chance, selection, and history. *Conservation Genetics*, **11**, 463-478.
- Brekke P, Bennett PM, Wang J, Pettorelli N, Ewen JG (2010) Sensitive males: inbreeding depression in an endangered bird. *Proceedings of the Royal Society B*, **277**, 3677-3684.
- Britt M, Haworth SE, Johnson JB, Martchenko D, Shafer ABA (2018) The importance of non-academic coauthors in bridging the conservation genetics gap. *Biological Conservation*, **218**, 118-123.
- Butchart HM, Walpole M, Collen B, *et al.* 2010. Global biodiversity: indicators of recent declines. *Science*, **328**, 1164-1168.
- Card DC, Schield DR, Reyes-Velasco J, Fujita MK, Andrew AL, Oyler-McCance SJ, Fike JA, Tomback DF, Ruggiero RP, Castoe (2014) Two low coverage bird genomes and a comparison of reference-guided versus *de novo* genome assemblies. *PLoS ONE*, **9**, e106649.
- Charlesworth D, Charlesworth B (1987) Inbreeding depression and its evolutionary consequences. *Annual Review of Ecology and Systematics*, **18**, 237-268.
- Crow JF, Kimura M (1970) *An introduction to population genetics theory*. Burgess Publishing Co., Minneapolis, Minnesota.
- Davis A (1994) Status, distribution, and population trends of the New Zealand shore plover *Thinornis novaeseelandiae*. *Notornis*, **41**, 179-194.
- de Villemereuil P, Rutschmann A, Lee KD, Ewen JG, Brekke P, Santure AW (2019) Little adaptive potential in a threatened passerine bird. *Current Biology*, **29**, 889-894.
- Dierickx, EG, Shulz AJ, Sato F, Hiraoka T, Edwards SV (2015) Morphological and genomic comparisons of Hawaiian and Japanese black-footed albatrosses (*Phoebastria nigripes*) using double digest RADseq: implications for conservation. *Evolutionary Applications*, **8(7)**, 662-678.
- Dodds KG, McEwan JC, Brauning R, Anderson RM, van Stijn TC, Kristansson T, Clarke SM (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, **16**, 1047.

- Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**(1), 51-63.
- Elliott GP, Merton DV, Jansen PW (2001) Intensive management of a critically endangered species: the kakapo. *Biological Conservation*, **99**, 121-133.
- Ewing SR, Nager RG, Nicoll MA, Aumjaud A, Jones CG, Keller LF (2008) Inbreeding and loss of genetic variation in a reintroduced population of Mauritis kestrel. *Conservation Biology*, **22**, 395-404.
- Fabian Y, Bollmann K, Brang P, Heiri C, Olschewski R, Rigling A, Stofer S, Holderegger (2019). How to close the science-practice gap in nature conservation? Information sources used by practitioners. *Biological Conservation*, **235**, 93-101.
- Fagan WF, Homes EE (2006) Quantifying the extinction vortex. *Ecology Letters*, **9**, 51-60.
- Faust LJ, Bergstrom YM, Thompson SD, Bier L (2018) PopLink version 2.4. Available at: <https://www.lpzoo.org/conservation-science/projects/poplink>.
- Fisch KM, Ivy JA, Burton RS, May B (2013) Evaluating the performance of conservation breeding techniques for conservation hatcheries: a case study of the delta smelt conservation breeding program. *Journal of Heredity*, **104**, 92-104.
- Forsdick NJ (*in prep.*) Minimal genomic impacts of hybridisation in kakī *Himantopus novaezelandiae* (Unpublished PhD Thesis). University of Otago, Dunedin, New Zealand.
- Forstmeier W, Schielzeth H, Mueller JC, Ellegren H, Kempenaers B (2012) Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. *Molecular Ecology*, **21**, 3237-3249.
- Frankham R (2005) Genetics and extinction. *Biological Conservation*, **126**, 131-140.
- Frankham R (2008) Genetic adaptation to captivity in species conservation programs. *Molecular Ecology*, **17**, 325-333.
- Frankham R (2010) Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, **143**, 1919-1927.
- Fraser DJ (2008) How well can conservation breeding programs conserve biodiversity? A review of salmonids. *Evolutionary Applications*, **1**, 535-586.
- Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations. *Molecular Ecology*, **26**, 5369-5406.

- Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology and Evolution*, **27**, 489-496.
- Galla SG, Buckley TR, Elshire R, *et al.* (2016) Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances. *Molecular Ecology*, **25(21)**, 5267-5281.
- Galla SJ, Forsdick NJ, Brown L, *et al.* (2019) Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management. *Genes*, **10**, 1-19.
- Garner BA, Hand BK, Amish SJ, Bernatchez L, Foster JT, Miller KM, Morin PA, Narum SR, O'Brien SJ, Roffler G, Templin WD, Sunnucks P, Strait J, Warheit KI, Seamons TR, Wenburg J, Olsen J, Luikart G (2017) Genomics in conservation: Case studies and bridging the gap between data and application. *Trends in Ecology and Evolution*, **31(2)**, 81-83.
- Glaser A, van Klink P, Elliott G, Edge K (2010) Whio/blue duck (*Hymenolaimus malacorhynchos*) recovery plan. Department of Conservation, Wellington: New Zealand.
- Goudet J, Kay T, Weir BS (2018) How to estimate kinship. *Molecular Ecology*, **27**, 4121-4135.
- Grueber CE, Jamieson IG (2008) Quantifying and managing the loss of genetic variation in a free-ranging population of takahe through the use of pedigrees. *Conservation Genetics*, **9**, 645-651.
- Grueber CE, Knafler GJ, King TM, Senior AM, Grosser S, Robertson B, Weston KA, Brekke P, Harris CLW, Jamieson IG (2015) Toll-like receptor diversity in 10 threatened bird species: relationship with microsatellite heterozygosity. *Conservation Genetics*, **16**, 595-611.
- Grueber CE, Reid-Wainscoat EE, Fox S, Belov K, Shier DM, Hogg CJ, Pemberton D (2017) Increasing generations in captivity is associated with increased vulnerability of Tasmanian devils to vehicle strike following release to the wild. *Scientific Reports*, **7**, 2161.
- Hagen EN, Hale ML, Maloney RF, Steeves TE (2011) Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the black stilt *Himantopus novaezelandiae*. *Ibis*, **153**, 556-561.
- Haig SM, Miller MP, Bellinger R, Draheim HM, Mercer DM, Mullins TD (2016) The conservation genetics juggling act: integrating genetics and ecology, science and policy. *Evolutionary Applications*, **9(1)**, 181-195.

- Hammerly SH, Morrow ME, Johnson JA (2013) A comparison of pedigree- and DNA-based measures for identifying inbreeding depression in the critically endangered Attwater's Prairie-chicken. *Molecular Ecology*, **22**, 5313-5328.
- Hammerly SC, de la Cerda DA, Bailey H, Johnson JA (2016) A pedigree gone bad: increased offspring survival after using DNA-based relatedness to minimize inbreeding in a captive population. *Animal Conservation*, **19(3)**, 296-303.
- Hayden EC (2014) Technology: the \$1,000 genome. *Nature*, **507**, 294–295.
- He X, Johansson ML, Heath DD (2016) Role of genomics and transcriptomics in selection of reintroduction source populations. *Conservation Biology*, **30(5)**, 1010-1018.
- Heber S, Briskie JV (2010) Population bottlenecks and increased hatching failure in endangered birds. *Conservation Biology*, **24**, 1674-1678.
- Hellmann JK, Sovic MG, Gibbs L, Reddon AR, O'Connor CM, Ligocki IY, Marsh-Rollo S, Balshine S, Hamilton IM (2016) Within-group relatedness is correlated with colony-level social structure and reproductive sharing in a social fish. *Molecular Ecology*, **25**, 4001-4013.
- Herrero-Medrano JM, Megens HJ, Crooijmans RP, Abellana JM, Ramis G (2012) Farm-by-farm analysis of microsatellite, mtDNA, and SNP genotype data reveals inbreeding and crossbreeding as threats to the survival of a native Spanish pig breed. *Animal Genetics*, **44(3)**, 259-266.
- Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, **93**, 47-64.
- Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, Thorne MAS, Lacy RC, Dasmahaptra KK (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Science*, **111**, 3775-3780.
- Hogg CJ, Grueber CE, Pemberton D, Fox S, Lee AV, Ivy JA, Belov K (2017) "Devil tools & tech": a synergy of conservation research and management practice. *Conservation Letters*, **10(1)**, 133-138.
- Hogg CJ, Taylor HR, Fox S, Grueber CE (2018) Response to Britt et al. 2018 "The importance of non-academic co-authors in bridging the conservation genetics gap". *Biological Conservation*, **218**, 118-123.
- Hogg CJ, Wright B, Morris KM, Lee AV, Ivy JA, Grueber CE, Belov K (2018) Founder relationships and conservation management: empirical kinships reveal the effect on breeding programmes when founders are assumed to be unrelated. *Animal Conservation*, **22(4)**, 348-361.

- Ivy JA, Miller A, Lacy RC, DeWoody J (2009) Methods and prospects for using molecular data in conservation breeding programs: an empirical example using Parma wallabies (*Macropus parma*). *Journal of Heredity*, **100**(4), 441-454.
- Ivy JA, Lacy RC (2012) A comparison of strategies for selecting breeding pairs to maximise genetic diversity retention in managed populations. *Journal of Heredity*, **103**, 186-196.
- Jamieson IG (2015) Significance of population genetics for managing small natural and reintroduced populations in New Zealand. *New Zealand Journal of Ecology*, **39**(1), 1-18.
- Johnson JA, Dunn PO. 2006. Low genetic variation in the Heath Hen prior to extinction and implications for the conservation of prairie-chicken populations. *Conservation Genetics* 7:37-48.
- Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**(10), 2511–2523.
- Jones OR, Wang J (2010) Molecular marker-based pedigrees for animal conservation biologists. *Animal Conservation*, **13**, 26-34.
- Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, **115**, 63-72.
- Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW (2016) genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, **9**, 1205-1218.
- Kearvell JC, Grant AD, Boon WM (2003) The orange-fronted parakeet (*Cyanoramphus malherbi*) is a distinct species: a review of recent research into its taxonomy and systematic relationship within the genus *Cyanoramphus*. *Notornis*, **50**, 27-35.
- Kearvell JC, Connor C, Farley M (2014) Field identification of the orange-fronted parakeet (*Cyanoramphus malherbi*): pitfalls for the unwary. *Notornis*, **61**, 200-204.
- Kearvell JC, Steeves TE (2015) Evidence for assortative mating in sympatric populations of orange-fronted (*Cyanoramphus malherbi*) and yellow-crowned (*C. auriceps*) kakariki. *Notornis*, **62**, 71–75.
- Kearvell JC, Legault AJ (2017) Orange-fronted parakeet. In New Zealand Birds Online. Retrieved from [www.nzbirdsonline.org.nz](http://www.nzbirdsonline.org.nz)
- Keedwell RJ, Maloney RF, Murray DP (2002) Predator control for protecting kaki (*Himantopus novaezelandiae*)— lessons from 20 years of management. *Biological Conservation*, **105**, 369-374.

- Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics*, **189**, 237-249.
- Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, Forstmeier W (2015) Quantifying realized inbreeding in wild and captive animal populations. *Heredity*, **114**, 397-403.
- Knight AT, Cowling RM, Rouget M, Balmford A, Lombard AT, Campbell BM (2008) Knowing but not doing: selecting priority conservation areas and the research-implementation gap. *Conservation Biology*, **22**, 610-617.
- Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. *Cell*, **155**, 27-38.
- Lacy RC (1995) Clarification of genetics terms and their use in the management of captive populations. *Zoo Biology*, **14(6)**, 565-577.
- Lacy RC (2009) Stopping evolution: genetic management of captive populations. In *Conservation genetics in the age of genomics*. Columbia University Press: 58-81.
- Lacy RC (2012) Extending pedigree analysis for uncertain parentage and diverse breeding system. *Journal of Heredity*, **103(2)**, 197-205.
- Lacy RC, Ballou JD, Pollak JP (2012) PMx: software package for demographic and genetic analysis and management of pedigreed populations. *Methods in Ecology and Evolution*, **3**, 433-437.
- Lemopoulos A, Prokkola JM, Uusi-Heikkilä A, Vasemagi A, Huusko A, Hyvärinen P, Koljonen M, Koskiniemi J, Vainikka A (2019) Comparing RADseq and microsatellites for estimating genetic diversity and relatedness – implications for brown trout conservation. *Ecology and Evolution*, **9(4)**, 1-19.
- Li M, Strandén I, Tiirikka T, Sevón-Aimonen M, Kantanen J (2011) A comparison of approaches to estimate the inbreeding coefficient and pairwise relatedness using genomic and pedigree data in a sheep population. *PLoS ONE*, **6(11)**, e26256.
- Li S, Li B, Cheng C, *et al.* (2014) Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biology*, **15**, 557.
- Maloney RF, Keedweel RJ, Wells NJ, Rebergen AL, Nilsson RJ (1999). Effect of willow removal on habitat use by five birds of braided rivers, Mackenzie Basin, New Zealand. *New Zealand Journal of Ecology*, **23**, 53-60.
- Marshall TC, Slate JBKE, Kruuk LEB, Pemberton JM (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular ecology*, **7(5)**, 639-655.



- McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Application of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526-538.
- Meffe GK, Carroll CR (1997) Principles of Conservation Biology (2nd ed.). Sinauer Associates Inc.: 3-27.
- Minias P, Dunn PO, Whittingham LA, Johnson JA, Oyler-McCance SJ (2019) Evaluation of a chicken 600K SNP genotyping array in non-model species of grouse. *Scientific Reports*, **9**, 6407.
- Muir P, Li S, Lou S, *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, **17**, 53.
- Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841-2847.
- Ortiz-Catedral L, Brunton DH (2009) Notes on the diet of the critically endangered orange-fronted parakeet (*Cyanoramphus malherbi*) on Maud Island. *New Zealand Journal of Zoology*, **36**, 385-388.
- Overbeek AL, Hauber ME, Brown E, Cleland S, Maloney RF, Steeves TE (2017) Evidence for brood parasitism in a critically endangered Charadriiform with implications for conservation. *Journal of Ornithology*, **158**(1), 333-337.
- Overbeek AL (2019) *Evidence for brood parasitism in a critically endangered Charadriiform with implications for conservation* (Unpublished Honours Thesis). Stanford University, Stanford, United States of America.
- Oyler-McCance SJ, Cornman RS, Jones KL, Fike JA (2015) Genomic single-nucleotide polymorphisms confirm that Gunnison and greater sage-grouse are genetically well differentiated and that the bi-state population is distinct. *Condor*, **117**(2), 217-227.
- Pemberton JM (2004) Measuring inbreeding depression in the wild: the old ways are the best. *Trends in Ecology and Evolution*, **19**(12), 613-615.
- Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society B*, **275**, 613-621.
- Pierce RJ (1983) The changed distribution of stilts in New Zealand. *Notornis*, **31**, 7-18.
- Pierce RJ (1996) Ecology and management of the black stilt (*Himantopus novaezelandiae*). *Bird Conservation International*, **6**(1), 81-88.

- Pierson JC, Coates DJ, Oostermeijer JGB, Beissinger SR, Bragg JG, Sunnucks P, Schumaker NH, Young AG (2016) Genetic factors in threatened species recovery plans on three continents. *Frontiers in Ecology and the Environment*, **14**(8), 433-440.
- Prince DJ, O'Rourke SM, Thompson TQ, Ali OA, Lyman HS, Saglam IK, Hotaling TJ, Spidle AP, Miller MR (2017) The evolutionary basis of premature migration in pacific salmon highlights the utility of genomics for informing conservation. *Science Advances*, **3**, e1603198.
- Putnam AS, Ivy JA (2014) Kinship-based management strategies for conservation breeding programs when pedigrees are unknown or uncertain. *Journal of Heredity*, **105**, 303-311.
- Ralls K, Ballou JD (1986) Captive breeding programs for populations with a small number of founders. *Trends in Ecology & Evolution*, **1**, 19-22.
- Rawlence NJ, Kennedy M, Scofield RP, Tennyson AJD, Boussès P, Chambers GK (2015) Taxonomic status of *Cyanoramphus* parakeets on the Auckland Islands and implications for the validity of the orange-fronted parakeet as a discrete species. *Journal of the Royal Society of New Zealand*, **45**, 197-211.
- Reed C (1998) Management plan for captive black stilts. *New Zealand Department of Conservation Threatened Species Occasional Publication*, **11**, 1-24.
- Robert A (2009) Captive breeding genetics and reintroduction success. *Biological Conservation*, **142**, 2915-2922.
- Robinson SP, Simmons LW, Kennington WJ (2013) Estimating relatedness and inbreeding using molecular markers and pedigrees: The effect of demographic history. *Molecular Ecology*, **22**, 5779-5792.
- Ross CT, WEise JA, Bon nar S, Nolin D, Trask JS, Smith DG, Ferguson B, Ha J, Kubisch HM, Vinson A, Kanthaswamy S (2014) An empirical comparison of short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) for relatedness estimation in Chinese rhesus macaques (*Macaca mulatta*). *American Journal of Primatology*, **76**, 313-324.
- Rudnick JA, Lacy RC (2008) The impact of assumptions about founder relationships on the effectiveness of conservation breeding strategies. *Conservation Genetics*, **9**, 1439-1450.
- Sanders MD, Maloney RF (2002) Causes of mortality at nests of ground-nesting birds in the Upper Waitaki Basin, South Island, New Zealand: a 5-year video study. *Biological Conservation*, **106**, 225-236.

- Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: Empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.
- Schulwitz S.E., Bedrosian B., and J.A. Johnson. 2014. Low neutral genetic diversity in an isolated greater sage grouse (*Centrocercus urophasianus*) population in northwest Wyoming. *The Condor* **116**, 560-573.
- Seddon PJ (2010) From reintroduction to assisted colonization: moving along the conservation translocation spectrum. *Restoration Ecology*, **18**, 796-802.
- Serrouya R Paetkau D, McLellan BN, Boutin S, Campbell M, Jenkins DA (2012) Population size and major valleys explain microsatellite variation better than taxonomic units for caribou in western Canada. *Molecular Ecology*, **21**, 2588-2601.
- Shafer ABA., Wolf JBW, Alves PC, *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, **30(2)**, 78-87.
- Silió L, Barragán C, Fernández AI, García-Casco J, Rodríguez MC (2015) Assessing effective population size, coancestry and inbreeding effects on litter size using the pedigree and SNP data in closed lines of the Iberian pig breed. *Journal of Animal Breeding and Genetics*, **133(2)**, 145-154.
- Slate J, David P, Dodds KG, Veenvliet BA, Glass BC, Broad TE, McEwan JC (2004) Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. *Heredity*, **93**, 255-265.
- Smith DHV, Converse SJ, Gibson KW, Moehrensclager A, Link WA, Olsen GH, Maguire K (2011) Decision analysis for conservation breeding: maximizing production for reintroduction of whooping cranes. *The Journal of Wildlife Management*, **75(3)**, 501-508.
- Smith DHV, Moehrensclager A, Christensen N, Knapik D, Gibson K, Converse SJ (2012) Archive eggs: a research and management tool for avian conservation breeding. *Wildlife Society Bulletin*, **36(2)**, 342-349.
- Snyder NFR, Derrickson SR, Beissinger SR, Wiley JW, Smith TB, Toone WD, Miller B (1996) Limitations of captive breeding in endangered species recovery. *Conservation Biology*, **10(2)**, 338-348.
- Soulé, M.E. 1985. What is conservation biology? A new synthetic discipline addresses the dynamics and problems of perturbed species, communities, and ecosystems. *Bioscience* **35(11)**, 727-734.
- Species360 (2017) Single Population Analysis and Record Keeping System (SPARKS). Available at: <http://www.vortex10.org/sparks16.html>

- Species360 (2019) Zoological Information Management System (ZIMS). Available at: [zims.Species360.org](https://zims.Species360.org).
- Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? *Nature*, **16**, 33-44.
- Steeves TE, Hale ML, Gemmell NJ (2008) Development of polymorphic microsatellite markers for the New Zealand black stilt (*Himantopus novaezelandiae*) and cross-amplification in the pied stilt (*Himantopus himantopus leucocephalus*). *Molecular Ecology Resources*, **8**, 1105-1107.
- Steeves TE, Maloney RF, Hale ML, Tylianakis JM, Gemmell NJ (2010) Genetic analyses reveal hybridization but no hybrid swarm in one of the world's rarest birds. *Molecular Ecology*, **19**, 5090-5100.
- Stoinski TS, Beck BB, Bloomsith MA, Maple TL (2003) A behavioral comparison of captive-born, reintroduced golden lion tamarins and their wild-born offspring. *Behavior*, **140(2)**, 137-160.
- Sunny A, Monroy-Vilchis O, Fajardo V, Aguilera-Reyes U (2014) Genetic diversity and structure of an endemic and critically endangered stream river salamander (Caudata: *Ambystoma leorae*) in Mexico. *Conservation Genetics*, **15**, 49-59.
- Taylor HR (2015) The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and Evolution*, **5(15)**, 3140-3150.
- Taylor HR, Kardos MD, Ramstad KM, Allendorf FW (2015) Valid estimates of individual inbreeding coefficients from marker-based pedigrees are not feasible in wild populations with low allelic diversity. *Conservation Genetics*, **16**, 901-913.
- Taylor HR, Soanes K (2016) Breaking out of the echo chamber: missed opportunities for genetics at conservation conferences. *Biodiversity and Conservation*, **25(10)**, 1987-1993.
- Taylor HR, Dussex N, van Heezik Y (2017) Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation*, **10**, 231-242.
- Thrasher DJ, Butcher BG, Campagna L, Webster MS, Lovette IJ (2018) Double-digest RAD sequencing outperforms microsatellite loci at assigning paternity and estimating relatedness: a proof of concept in a highly promiscuous bird. *Molecular Ecology Resources*, **18**, 953-965.
- Tzika AC, Remy C, Gibson R, Milinkovitch MC (2009) Molecular genetic analysis of a captive-breeding program: the vulnerable endemic Jamaican yellow boa. *Conservation Genetics*, **10**, 69-77.

- van Heezik Y, Lei P, Maloney R, Sancha E (2005) Captive breeding for reintroduction: influence of management practices and biological factors on survival of captive kaki (black stilt). *Zoo Biology*, **24**, 459-474.
- Walters JR, Derrickson SR, Fry DM, Haig SM, Marzluff HM, Wunderle JM (2010) Status of the California condor (*Gymnogyps californianus*) and efforts to achieve its recovery. *The Auk*, **127(4)**, 969-1001.
- Wan QH, Wu H, Fujihara T, Fang SG (2004) Which genetic markers for which conservation genetic issue? *Electrophoresis*, **25**, 2165-2176.
- Wang J (2011) COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, **11(1)**, 141-145.
- Wang J (2013) A simulation module in the computer program COLONY for sibship and parentage analysis. *Molecular Ecology Resources*, **13(4)**, 734-739.
- Wang J (2016) Pedigrees or markers: which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, **107**, 4-13.
- Weir BS, Anderson AD, Hepler AB (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, **7**, 771-780.
- Wilcove DS, Rothstein D, Dubow J, Phillips A, Losos E (1998) Threats to imperilled species in the United States. *BioScience*, **48(8)**, 607-615.
- Williams SE, Hoffman EA (2009) Minimizing genetic adaptation in captive breeding programs: a review. *Biological Conservation*, **142**, 2388-2400.
- Willoughby JA, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, Dewoody A (2015) The impacts of inbreeding, drift and selection on genetic diversity in conservation breeding populations. *Molecular Ecology*, **24**, 98-110.
- Wright S (1922) Coefficients of inbreeding and relationship. *The American Naturalist*, **56**, 330-338. <http://cantm1.canterbury.ac.nz/printers/>
- Wright B, Morris K, Grueber CE, Willet CE, Gooley R, Hogg C, O'Meally D, Hamede R, Jones M, Wade C, Belov K (2015) Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population. *BMC Genomics*, **16**, 791-802.
- Woolaver LG (2013) Population genetics and relatedness in a critically endangered island raptor, Ridgeway's hawk *Buteo ridgwayi*. *Conservation Genetics*, **14**, 559-571.

## Chapter 2: Pedigree Development and Investigation for Kākāriki Karaka and Kakī Conservation

### 2.1: Preamble

This chapter consists of two reports for the Department of Conservation — in between a brief General Introduction and Concluding Remarks— detailing the development of pedigree resources for kākāriki karaka and kakī. The report for kākāriki karaka was written in July 2018, whereas the report for kakī was written in July 2019. As discussed in **Chapter 1**, conservation breeding is an approach that allows practitioners to prevent species extinction and enhance recovery. In these programmes, pedigrees are commonly used to make pairing decisions that minimise inbreeding and maximise diversity (Ballou *et al.* 2010). The pedigrees for kākāriki karaka and kakī have never been formally investigated, and both the Kākāriki Karaka and Kakī Recovery Groups have questions that could be informed by pedigrees for conservation management. In **Chapter 2**, we aim to address these questions using multigenerational pedigrees developed for both species. This chapter reflects upon limitations of these pedigrees and how empirical estimates of relatedness may be better suited to inform particular management actions. With these caveats in mind, the pedigree resources developed in **Chapter 2** are also integral to research conducted in **Chapter 5** and will be invaluable for future research and management (**Chapter 6**).

#### 2.1.1: Contribution Statement

I conceptualised this chapter in response to the need for pedigrees in kākāriki karaka and kakī conservation practice (**Chapter 2**) and research (**Chapters 5 and 6**). I led the construction of the kakī studbook using 40 years of data collected by dozens of staff from

the Department of Conservation. The kākārīki karaka studbook used in this chapter was largely constructed by Anne Richardson (Wildlife Manager, Isaac Conservation and Wildlife Trust; also Captive Coordinator for the Kākārīki Karaka Recovery Group) and I updated it with input from her. Both studbooks were corrected for errors during pedigree development, in consultation with conservation practitioners from the Department of Conservation (i.e., Liz Brown, Simone Cleland, Cody Thyne) and The Isaac Conservation and Wildlife Trust (i.e., Anne Richardson). I performed all pedigree analyses and designed all figures and tables presented in this chapter. I wrote all drafts of this chapter, with feedback from Tammy Steeves, Anna Santure, Anne Richardson, and Liz Brown.



**Stephanie J. Galla**

## 2.2: Department of Conservation Pedigree Reports for Kākārīki Karaka and Kakī

### 2.2.1: Abstract

Pedigrees are a common tool in the genetic management of threatened species with conservation breeding programmes. However, these resources have never been developed or formally investigated for the conservation of the critically endangered kākārīki karaka (*Cyanoramphus malherbi*) or kakī (*Himantopus novaezelandiae*). This chapter includes two reports for the Department of Conservation that provide pedigree resources for kākārīki karaka (written in July 2018) and kakī (written in July 2019) and a brief analysis of founder representation, inbreeding, pairwise relatedness, and mean kinship. This chapter provides new pedigree resources for two critically endangered birds endemic to Aotearoa New

Zealand. This chapter also underscores the benefit of generating empirical estimates of relatedness to augment pedigrees, particularly for individuals of unknown parentage used to supplement conservation breeding programmes.

### 2.2.2: General Introduction

Pedigrees are a fundamental tool in the fields of ecology, evolution, and conservation, allowing researchers to gain insight into ancestry, selection, trait heritability, and behaviour (Pemberton 2008; Jones & Wang 2010). In a conservation context, multigenerational pedigrees have been important for the demographic and genetic management of threatened species within conservation breeding programmes. In these programmes, individuals are strategically paired and bred in intensively managed landscapes (i.e., captivity, island or mainland sanctuaries) to enhance recovery efforts (Ballou *et al.* 2010). Strategic pairings are made using pedigree data to simultaneously minimise inbreeding and prioritise individuals with low mean kinship (i.e., the average relatedness between a non-founding individual to all others in a population, including oneself; Lacy 1995; Ballou & Lacy 1995; Ivy & Lacy 2012). Offspring from these pairings may remain in captivity as an insurance population (as suggested for some captive kea, *Nestor notabilis*; Orr-Walker *et al.* 2015), remain on intensively managed island or mainland sanctuaries (e.g., kākāpō, *Strigops habroptilus*, Elliott *et al.* 2001; takahē, *Porphyrio hochstetteri*, Grueber & Jamieson 2008), or they may be translocated to supplement wild populations (e.g., whio/blue duck, *Hymenolaimus malacorhynchos*, Glaser *et al.* 2010). The goal of this approach is to prevent extinction and maximise genetic diversity in an effort to enhance species resilience to environmental change (i.e., evolutionary potential; de



Villemereuil *et al.* 2019; Giglio *et al.* 2016). To date, many conservation practitioners have used a conservation breeding approach for the demographic and genetic management of threatened populations, including over 350 species worldwide that are captive bred for translocation to the wild (Smith *et al.* 2011).

Pedigrees are often considered the tool of choice for informing genetic management in conservation breeding programmes, as they are based on long-standing principles of genome sharing by Mendelian inheritance (Wright 1922). Pedigree-based management approaches are also supported by extensive research demonstrating how genetic diversity can be maintained through minimising pedigree-based mean kinship and pairwise inbreeding (Ballou 1983; Ballou & Lacy 1995; Rudnick & Lacy 2008; Lacy 2009; Ballou *et al.* 2010; Ivy & Lacy 2012; Putnam & Ivy 2014; Willoughby *et al.* 2015). Software programmes are readily available software to maintain ancestry databases (i.e., studbooks; *PopLink*, Faust *et al.* 2012; *SPARKS*, Species360 2017; *ZIMS*, Species360 2019) and analyse pedigrees (e.g., *PMx*, Lacy *et al.* 2012), making pedigree-based approaches practical for the genetic management of species with conservation breeding programmes.

While pedigrees are often used for conservation genetic management, there are caveats that may hamper pedigree accuracy. For example, all pedigrees assume the founders of the population are unrelated (Ballou 1983), which is unlikely in threatened species where founders are typically sourced from small populations that have experienced one or more sustained demographic bottlenecks (Hogg *et al.* 2018). This assumption can be repeatedly violated when individuals of unknown parentage are periodically added into the population (Ivy *et al.* 2016). Research indicates that pedigrees of substantial depth ( $\geq$  five generations recorded) are robust enough to

accurately estimate kinship and inbreeding despite this violation (Balloux *et al.* 2004; Pemberton 2004; Rudnick & Lacy 2008). However, conservation breeding programmes that are routinely augmented with individuals of unknown ancestry may have perpetually shallow pedigrees that are susceptible to underestimated kinship and inbreeding coefficients (Russello & Amato 2004; Ivy *et al.* 2016; Hogg *et al.* 2018). Pedigrees may also be hindered because of data entry errors, which can compound over many generations (Hammerly *et al.* 2016). Even with these caveats in mind, pedigrees remain a powerful tool that have been crucial for the genetic and demographic management of species on the brink of extinction around the globe (Ballou *et al.* 2010).

In Aotearoa New Zealand, there are over 20 threatened species with conservation breeding programmes (Lucy Bridgeman, *personal comm.*; see also Miskelly & Powlesland 2013), including programmes for kākāriki karaka/orange-fronted parakeet (*Cyanoramphus malherbi*; Ortiz-Catedral *et al.* 2009) and kakī/black stilt (*Himantopus novaezelandiae*; van Heezik *et al.* 2005). Parentage data has been collected for both of these critically endangered species for decades. However, the use of pedigrees for management in these two birds has been limited. The aim of the following reports is to develop pedigree resources for kākāriki karaka and kakī, and to demonstrate their utility for answering questions of interest to the recovery groups for both species.

### **2.2.3: Kākāriki Karaka Pedigree Report, July 2018**

*This report was prepared for the Kākāriki Karaka Recovery Group for during their annual meeting in July 2018. It has since been edited for clarity to be included in this thesis*

*chapter. House names for captive birds (e.g., “Maverick”, “Poldark”) have been used in the text for ease of interpretation by conservation practitioners.*

### **2.2.3.1: Kākāriki Karaka Conservation Breeding History**

The conservation breeding programme for kākāriki karaka was initiated in 2003, after this species experienced significant decline in the wild following two consecutive years of beech forest (*Nothofagus* spp.) mast-induced rat plagues (BirdLife International 2018). Founding individuals were originally sourced from three valleys in north Canterbury: the south branch of the Hurunui (Hurunui, hereafter; years sourced: 2003, 2004), the Hawdon (years sourced: 2004, 2011, 2014), and Poulter (years sourced: 2009, 2016; Department of Conservation, *personal comm.*). Captive pairs are currently bred at the Isaac Conservation and Wildlife Trust (Christchurch, Aotearoa), Orana Wildlife Park (Christchurch, Aotearoa), and the Auckland Zoo (Auckland, Aotearoa), with an additional pair currently being established at Pūkaha National Wildlife Centre (Tararua, Aotearoa; Anne Richardson, *personal comm.*). Historically, offspring from these pairings have been released to four predator free offshore islands, including Oruawairua/Blumine Island, Te Pākeka/Maud Island, Te Kāhahu-o-Tamatea/Chalky Island, and Tuhua/Mayor Island. More recently, birds have been released to a predator-controlled area in the Hurunui. Currently, the Hurunui is the only release site for captive-reared kākāriki karaka. Oruawairua and the Hurunui are the only contemporary locations with established breeding populations (Department of Conservation, *personal comm.*; Figure 2.1).

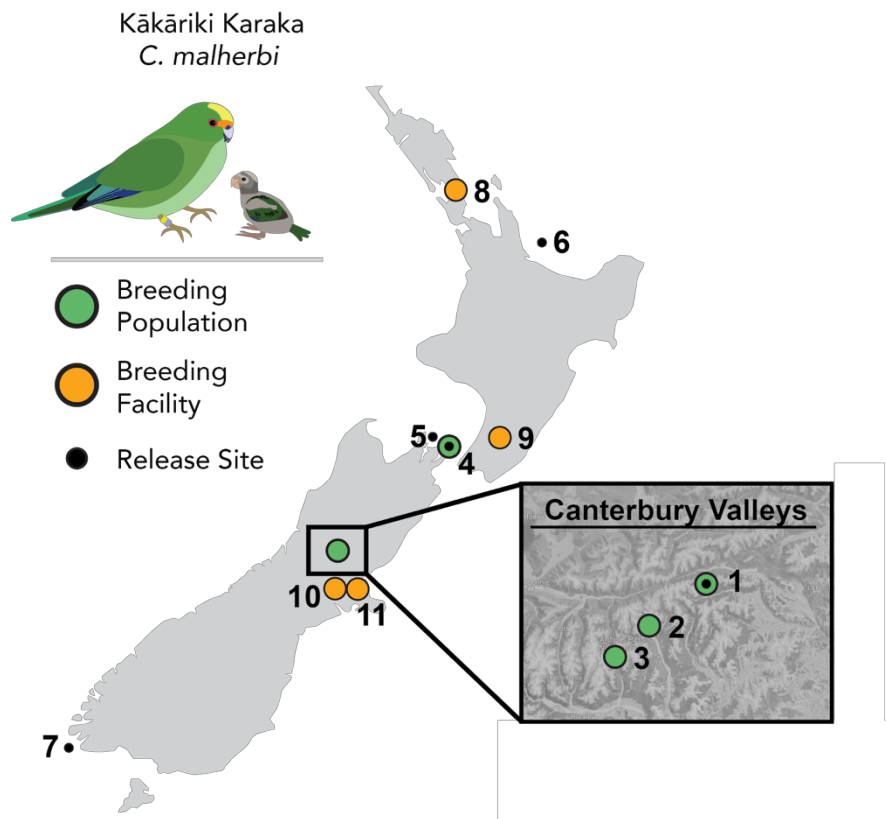


Figure 2. 1: Map detailing kākāriki karaka breeding populations: 1 — Hurunui, 2 — Poulter Valley, 3 — Hawdon Valley, 4 — Oruawairua/Blumine Island; release sites: 1 — Hurunui, 4 — Oruawairua/Blumine Island, 5 — Te Pākeka/Maud Island, 6 — Tuhua/Mayor Island, 7 — Te Kākuha-o-Tamatea/Chalky Island; and current captive breeding facilities: 8 — Auckland Zoo, 9 — Pūkaha Wildlife Centre, 10 — Orana Wildlife Park, 11 — Isaac Conservation and Wildlife Trust). Tuhua, Te Pākeka, and Te Kākuha-o-Tamatea do not currently sustain breeding populations. Pūkaha Wildlife Centre currently holds individuals available for breeding, but at the time of writing of this thesis, no active breeding pairs.

Life history data (e.g., hatch date, parentage, sex, relevant translocations, and status) have been collected for all captive birds since 2003 and have been maintained in the studbook management programme SPARKS (Species360 2017). In lieu of making pairing recommendations based on pedigree-based kinship values, captive pairs have been formed to generally avoid inbreeding and equalise founder representation by evaluating visualised pedigree diagrams (Anne Richardson, *personal comm.*). While there is a microsatellite panel developed for kākāriki karaka (Andrews *et al.* 2013), these markers have not been used to assess relatedness between potential pairs. In the wild, kākāriki karaka can hybridise with another endemic congener, the yellow-crowned parakeet (*Cyanoramphus auriceps*; Kearvell

& Grant 2003), but hybrids have been actively excluded from the conservation breeding programme.

Currently, kākāriki karaka have likely been extirpated from the Hawdon Valley, and few individuals remain in the Poulter Valley (Department of Conservation, *personal comm.*). Given the high potential for additional rat plagues —for example, following the mega mast forecast for 2019— the Department of Conservation has recently prioritised the formation of more captive pairs (Department of Conservation, *personal comm.*). To this end, kākāriki karaka eggs of wild parentage were harvested from one clutch in the Poulter Valley of North Canterbury during the 2016-2017 breeding season, with one individual (House Name: 'Poldark') surviving to breeding age and incorporated into the breeding programme. In addition to Poldark, the Department of Conservation has considered sourcing individuals from Oruawairua, a former release site for captive reared birds from 2011-2012. This source may be beneficial to captive management, as there may be founders represented by individuals released to Oruawairua that are no longer represented in the captive population. While the pedigree has never been formally analysed for kākāriki karaka, there is an opportunity to use pedigree resources developed for this species to provide an indication of whether founder lineages found on Oruawairua may be different from those found in captivity. Additionally, available pedigree data could be used to estimate relatedness, inbreeding, and mean kinship among potential breeders —including Poldark— to make pairing recommendations for the conservation breeding programme.

### *2.2.3.2: Kākāriki Karaka Pedigree Development*

Life history information (e.g., hatch date, parentage, sex, relevant translocations, and status) was migrated from the existing SPARKS studbook to PopLink v. 2.5.1 (Faust *et al.*

2018), an open source and user-friendly software for maintaining and editing studbooks. The kākāriki karaka studbook was manually updated with new breeding pairs and offspring from the 2016-2017 and 2017-2018 breeding seasons ( $n = 624$  birds total). The studbook was examined for data entry errors (e.g., misentered sex, status, or parentage) using the *Validation Report* tool within PopLink. Many errors were systematically resolved in collaboration with Anne Richardson at the Isaac Conservation and Wildlife Trust.

Pedigree analyses were performed using the programme PMx v. 1.6.20190628 (Lacy *et al.* 2012). Pedigree descriptive statistics were calculated while selecting for all known individuals, including sex ratio, age range, gene diversity, number of founders, founder genome equivalents, mean inbreeding, average mean kinship, mean generation time, percent known ancestry, and effective population size. In addition to these statistics using all individuals, founder representation was calculated for the birds in Oruawairua ( $n = 62$ ) and those in captivity ( $n = 14$ ) to determine differences in founder representation.

A pool of breeding individuals ( $n = 37$ ) are available for captive pairing. These individuals include offspring from captive pairs Gabby & Mavrick ( $n = 7$ ), Storm & Abbie ( $n = 7$ ), Hone & Daisy ( $n = 10$ ), Tom & Minie ( $n = 8$ ), Tama & Green ( $n = 2$ ) and Greenie & Lory ( $n = 2$ ), hatched during the 2017-2018 breeding season. Poldark, an individual of unknown wild parentage from the Poulter Valley, is also available for pairing. Founder representation, pairwise kinship, inbreeding coefficients, and mean kinship values were produced for all 37 potential breeding individuals. Pairwise relatedness values between these birds were calculated using the formula  $R(xy) = 2 * f(xy) / \sqrt{(1+F_x)(1+F_y)}$ . In this formula,  $f(xy)$  is the kinship between two individuals ( $x$  and  $y$ ) and  $F_x$  and  $F_y$  are the inbreeding coefficients of individuals  $x$  and  $y$  (Crow & Kimura, 1970).

### 2.2.3.2: Kākāriki Karaka Pedigree and Founder Representation

The entirety of the kākāriki karaka pedigree as of January 2018 includes 14 founders, or 5.9 founder genome equivalents (Table 2.1). The oldest recorded kākāriki in captivity was 16.4 years old and the sex ratio of individuals in the pedigree is equal. Gene diversity is high ( $> 0.2$ ), and mean inbreeding (0.03) and average mean kinship (0.09) are low (Table 2.1). There are no individuals of unknown parentage in the kākāriki karaka pedigree, but it is relatively shallow (maximum 4.5 generations deep, average depth = 2.45 across all individuals; Table 2.1).

Table 2. 1: Pedigree descriptive statistics for all individuals in the kākāriki karaka pedigree, as produced by PMx.

Diversity Statistic	All Individuals
# Individuals	624
Sex Ratio (% Males)	0.50
Age Range	<16.4 Years
Gene Diversity	0.92
# Founders	14
Founder Genome Equivalents	5.90
Mean Inbreeding	0.03
Average MK	0.09
Mean Generation Time	1.36
% Ancestry Known	100
% Ancestry Certain	100
Ne/N	0.07

The current captive population has 12 of the 14 founders represented, with an average pedigree depth of 2.32 generations. The two lineages that are not represented in the captive population belong to individuals WILD\_A and WILD\_B, both sourced from the Hurunui. These two birds only produced one fledgling, which was released to Te Pākeka. In the captive population (i.e., breeding individuals and offspring available for pairing), the proportion of founder representation from the Hawdon and Hurunui Valleys is higher than representation from the Poulter Valley (Table 2.2).

Table 2. 2: Founder representation across all individuals in the population, captive breeders, captive offspring currently available for pairing (except Poldark), and individuals released to Oruawairua in 2011-2012.

Unique ID	Valley Origin	Year Sourced	All Individuals	Founder Representation		
				Captive Breeders	Captive Offspring	Oruawairua
Wild A	Hurunui	2003	0.001	0	0	0
Wild B	Hurunui	2003	0.001	0	0	0
Wild 1	Hurunui	2003	0.0975	0.0692	0.0515	0.0887
Wild 2	Hurunui	2003	0.0975	0.0692	0.0515	0.0887
Wild 5	Hurunui	2004	0.0882	0.1563	0.1618	0.0806
Wild 6	Hurunui	2004	0.0882	0.1563	0.1618	0.0806
<b>Hurunui Founders Combined</b>			<b>0.3734</b>	<b>0.451</b>	<b>0.4266</b>	<b>0.3386</b>
Wild 7	Poulter	2009	0.0958	0.0759	0.0662	0.1492
Wild 8	Poulter	2009	0.0958	0.0759	0.0662	0.1492
<b>Poulter Founders Combined</b>			<b>0.1916</b>	<b>0.1518</b>	<b>0.1324</b>	<b>0.2984</b>
Wild 3	Hawdon	2004	0.1346	0.0692	0.0515	0.1169
Wild 4	Hawdon	2004	0.1346	0.0692	0.0515	0.1169
Wild 9	Hawdon	2011	0.042	0.0402	0.0368	0.0645
Wild 10	Hawdon	2011	0.042	0.0402	0.0368	0.0645
Wild 11	Hawdon	2014	0.0334	0.0893	0.1324	0
Wild 12	Hawdon	2014	0.0334	0.0893	0.1324	0
<b>Hawdon Founders Combined</b>			<b>0.42</b>	<b>0.3974</b>	<b>0.4414</b>	<b>0.3628</b>

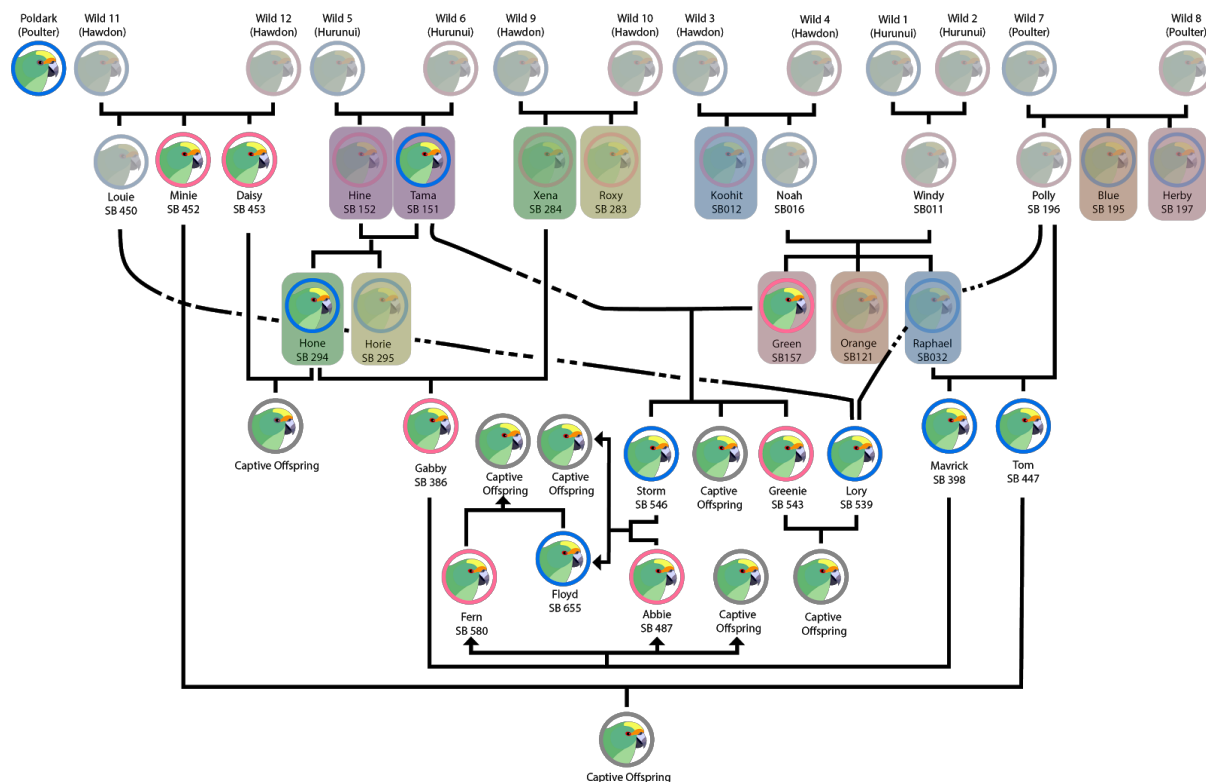


Figure 2. 2: Visualised diagram of the kākāriki karaka captive pedigree. Black lines denote captive pairs that have produced current captive offspring available for pairing. Coloured boxes behind individuals denote captive pairs that produced offspring released on Oruawairua from 2011-12. Faded birds denote dead individuals.



Birds released to Oruawairua were offspring from six breeding pairs, representing 10 of the 14 founders. All founders represented in Oruawairua are still represented in the current captive population (Table 2.2, Figure 2.2). While the same lineages found in Oruawairua can be found in captivity, the proportion of founder contribution differs between populations. For example, the proportion of founder representation from the Poulter Valley (0.3) in individuals released to Oruawairua is double what is found in captivity (0.15).

#### 2.3.3.4: Kākāriki Karaka Pairing Recommendations

Pedigree depth for individuals available for pairing is low (on average 2.25 generations deep). Inbreeding coefficients of these individuals are zero, except for the offspring of Storm & Abbie, who have some shared ancestry. Mean kinship is non-zero for almost every bird, with the exception of Poldark, who is of wild ancestry and assumed to be unrelated to all others in the population (Table 2.3).

*Table 2. 3: Individual genetic details of potential breeding individuals, including inbreeding coefficient (F), mean kinship (MK), and pedigree depth. n denotes the number of full siblings that share the same genetic details for this breeding season. It should be noted that MK fluctuates depending on the pool of individuals analysed within PMx.*

<i>n</i>	Parents	<i>F</i>	<i>MK</i>	Pedigree Depth
1	Wild (Poldark)	0	0	0
8	Tom & Minie	0	0.1072	2.75
7	Storm & Abbie	0.0781	0.1439	4
7	Gabby & Mavrick	0	0.1313	3.5
10	Hone & Daisy	0	0.1395	2.5
2	Tama & Green	0	0.1079	2.5
2	Greenie & Lory	0	0.0995	3.25

Pairwise relatedness (*R*) estimates between all potential breeders indicate that most pairings will result in some degree of inbreeding, with the exception of Poldark, who appears unrelated to all other individuals in the captive population (Tables 2.3 and 2.4). As a

consequence, amongst all individuals available for pairing, pedigree-based mean kinship and pairwise relatedness estimates suggest that Poldark is suitable to mate with anyone. Beyond Poldark, these results suggest that offspring of Tom & Minie and Tama & Green would be suitable for pairing, as they have lower mean kinship values and relatively low pairwise relatedness values with each other (Tables 2.3 and 2.4).

*Table 2. 4: Pedigree-based estimates of relatedness (R) between all potential breeders. Individuals highlighted in red have first order relatedness ( $R > 0.5$ , pairing not recommended), orange have greater than half-sibling or double first cousin relatedness ( $R = 0.25-0.49$ , pairing not recommended), pale yellow have greater than first cousin relatedness ( $R = 0.125-0.245$ , pairing not recommended), pale green have greater than first cousin once removed relatedness ( $R = 0.0625-0.125$ , can be paired), and grass green are relatively unrelated ( $R < 0.0625$ , ideal pairing).*

SB #		603	604	606	610	613	614	616	618	619	620	622	624	630	631	1001	1003
Parents		Abby & Storm	Abby & Storm	Hone & Daisy	Tom & Minie	Tom & Minie	Tom & Minie	Abby & Storm	Abby & Storm	Gabby & Maverick	Gabby & Maverick	Gabby & Maverick	Hone & Daisy	Hone & Daisy	Hone & Daisy	Tama & Green	Greenie & Lory
Sex		F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
578 Gabby & Mav	M																
579 Gabby & Mav	M																
584 Poldark (Wild)	M																
586 Abby & Storm	M																
605 Hone & Daisy	M																
607 Tom & Minie	M																
608 Tom & Minie	M																
609 Tom & Minie	M																
611 Tom & Minie	M																
612 Tom & Minie	M																
615 Abby & Storm	M																
617 Abby & Storm	M																
621 Gabby & Mav	M																
623 Gabby & Mav	M																
625 Hone & Daisy	M																
626 Hone & Daisy	M																
627 Hone & Daisy	M																
628 Hone & Daisy	M																
629 Hone & Daisy	M																
1000 Tama & Green	M																
1002 Greenie & Lory	M																

### 2.3.3.5: Kākāriki Karaka Pedigree Discussion

A kākāriki karaka pedigree has been developed in collaboration with the Department of Conservation and the Isaac Wildlife and Conservation Trust, including 624 captive individuals from 2003-present. Founder representation analyses indicate that two founding lineages have been lost since the time the conservation breeding programme was initiated and 12 founding lineages remain in captivity. Pedigrees operate under the assumption that founders are unrelated (Ballou 1983). However, founders from conservation breeding

programmes are often sourced from small populations after one or more demographic bottlenecks, resulting in individuals that may be related to one another (e.g., Bergner *et al.* 2014). Pedigrees that are deep (i.e.,  $\geq$  five generations recorded) are more robust to this assumption and provide more accurate measures of kinship and inbreeding (Balloux *et al.* 2004; Pemberton *et al.* 2008; Rudnick & Lacy 2008). Like many other conservation breeding programmes where individuals of unknown parentage are routinely sourced (e.g., Ivy *et al.* 2016; Hogg *et al.* 2018), the kākāriki karaka pedigree is relatively shallow, with many founders having joined the population in recent years. The results given here should be interpreted with caution, as founders of the captive kākāriki karaka population may be inadvertently related to one another, leading to underestimated kinship and inbreeding values (Russello & Amato 2004).

With this caveat in mind, the pedigree developed here provides helpful ancestry information regarding individuals released to Oruawairua. There are no lineages represented in the birds that were released on Oruawairua that are not captured by individuals in captivity. Still, there are subtle differences in the proportion of founder representation, with a higher proportion of Poulter Valley founders represented by the individuals released to Oruawairua. These differences demonstrate how the disproportionate production of chicks by a few breeding pairs can change founder representation over time. It is also likely that the genetic composition of both the captive and Oruawairua populations has changed over time due to genetic drift, as has been documented in other translocated bird populations in Aotearoa (e.g., hihi, *Notiomystis cincta*, Brekke *et al.* 2011; black robin, *Petroica traversi*, Forsdick *et al.* 2017). This suggests that some alleles currently found on Oruawairua may not be represented in captivity, and vice versa. We recommend sourcing clutches from Oruawairua to augment the captive

breeding population during the 2018-2019 breeding season. Prior to flock mating in captivity, we also recommend using empirical estimates of kinship (i.e., pairwise relatedness calculated from genome-wide single nucleotide polymorphisms, or SNPs; see **Chapters 5 and 6**) for any birds sourced from Oruawairua to determine relatedness between these birds of unknown parentage and those in the captive population.

Pedigree-based estimates of relatedness, inbreeding, and mean kinship indicate that Poldark, a wild individual sourced from the Poulter Valley, is suitable to pair with any other individual in the population. However, these recommendations are based on the assumption that Poldark, as a founder, is unrelated to all other individuals in the population. Given that the Poulter Valley of North Canterbury has had relatively few individuals in the population over the last two decades, it is unlikely that Poldark and other Poulter Valley founders (i.e., Wild 7 and Wild 8) are unrelated to one another. Genetic studies to date show weak genetic differentiation between individuals from these Canterbury valleys (Andrews 2013), but the possibility of relatedness between any founders from different valleys cannot be precluded. Pedigree-based relatedness and mean kinship values also suggest that offspring from Tom & Minie and Tama & Green would make suitable pairings. With that said, these recommendations are also based on relatively shallow pedigrees for these individuals. Given the low pedigree depth in the kākāriki karaka pedigree, we recommend the inclusion of empirical estimates of relatedness using genome-wide SNPs (Goudet *et al.* 2018) to improve the accuracy and utility of the kākāriki karaka pedigree in the years to come (see **Chapters 5 and 6**).

#### *2.3.4: Kakī Pedigree Report, August 2019*

*This report has been prepared for the Kakī Recovery Group and will be presented at the next meeting in 2020, with recommendations shared with conservation practitioners in the interim.*

##### *2.3.4.1: Kakī Conservation Breeding History*

The kakī conservation breeding programme was formally initiated in the late 1970's after the population experienced significant population decline due to introduced mammalian predators and habitat loss (Sanders & Maloney 2002). This breeding programme includes two components: a wild rearing component where eggs are harvested from intensively monitored wild pairs and artificially reared in captivity, and a captive breeding component where pairs are bred in captivity (Maloney & Murray 2001; van Heezik *et al.* 2005). Captive breeding was initially started at Pūkaha National Wildlife Centre. Today, captive breeding, artificial incubation, and hand rearing takes place at the Department of Conservation Kakī Captive Management Centre in Twizel and the Isaac Conservation and Wildlife Trust in Christchurch. When offspring from these pairs are juvenile or subadult age, they are translocated to predator controlled sites in Te Manahuna/The Mackenzie Basin, including most recently a site at Glentanner near the Tasman River Delta, and another at Mt. Gerald near the Godley River delta (Figure 2.3).

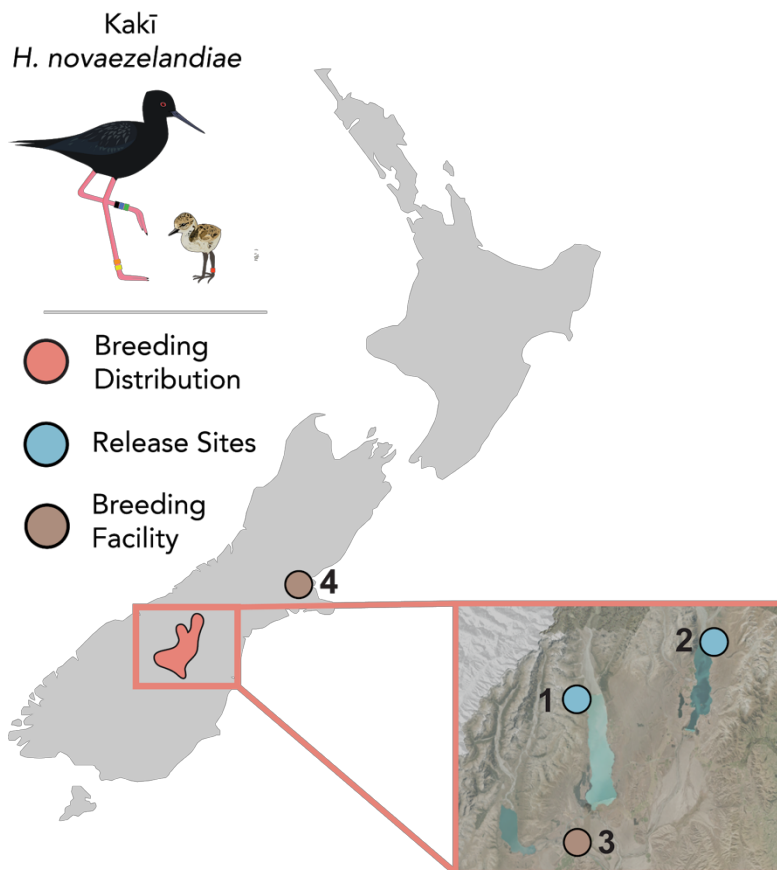


Figure 2. 3: Map detailing kakī breeding range, release sites (1 — Glentanner, 2 — Mt. Gerald), and captive breeding/rearing facilities (3 — DOC Kakī Captive Management Centre, 4 — Isaac Conservation and Wildlife Trust) in Aotearoa.

Life history data (e.g., hatch date, parentage, sex, relevant translocations, and status) has been collected for managed individuals since the late 1970's, and is kept in a Department of Conservation spreadsheet entitled the "alpha list". A related spreadsheet entitled "egg chick" is also maintained, which additionally includes eggs that are infertile, addled, suffer mortality during incubation, and chicks that do not survive to banding age. In addition to these documents, hand drawn pedigree diagrams for all individuals are maintained by the Department of Conservation (Department of Conservation, *personal comm.*; Figure 2.4).



reduced reproductive success (Wallis 1999; Steeves *et al.* 2010), which may contribute to the lack of introgression seen in node J kakī markers (Steeves *et al.* 2010; Forsdick *et al. in prep.*).

In addition to demographic considerations, captive pairing recommendations have also been made to generally avoid inbreeding between pairs. These pairing recommendations often begin with a list of candidates from the captive or wild rearing programme that appear relatively unrelated and representative of founder diversity based on visualised pedigree diagrams. To ensure that the candidate breeding individuals identified using these diagrams are indeed unrelated, from 2010 the Kakī Recovery Programme began using eight microsatellite markers developed for kakī (Steeves *et al.* 2008) to estimate relatedness between individuals. Individuals with the lowest pairwise relatedness are prioritised for pairing, to avoid negative fitness consequences associated with inbreeding (Hagen *et al.* 2011, but see also **Chapter 5**).

While avoiding inbreeding is important for genetic management, an additional consideration is the retention and equalisation of founder diversity, which can be captured using a mean kinship strategy (Ballou & Lacy 1995). To this end, conservation practitioners in the Kakī Recovery Group are interested in learning whether certain founders are over-represented in the population, and whether individuals should be given priority for future breeding strategies given their mean kinship values. To date, generating kinship values using kakī ancestry data has been limited, as it is difficult to perform pedigree analyses with spreadsheets or paper documents. However, generating a kakī pedigree is possible given the long-term data sets available from the Department of Conservation, and presents an exciting opportunity to develop a resource for kakī research and recovery efforts. Here, we aim to develop pedigree



resources for kakī and explore founder representation and mean kinship. Through this endeavour, we aim to identify whether overrepresented lineages exist and whether there are individuals that should be prioritised for future pairings.

#### *2.3.4.2: Kakī Pedigree Construction*

A multigenerational pedigree was constructed for kakī by entering studbook information (i.e., hatch date, sex, parentage, and status) for every managed bird since the late 1970's ( $n = 2,680$  birds) into the studbook management software PopLink v. 2.5.1 (Faust *et al.* 2018). The “black stilt” or “BS” number was used as the unique identifier (or studbook number) for each individual. This studbook was examined for errors (e.g., misentered sex, status, or incorrect parentage given other life history information) using the *Validation Report* tool within PopLink. Many of these errors were resolved in collaboration with Liz Brown, Simone Cleland, and Cody Thyne at the Department of Conservation.

Pedigree analyses were performed using the programme PMx v. 1.6.20190628 (Lacy *et al.* 2012), with all unknown individuals in the pedigree treated as wild founders. To understand how many founders have existed in the programme, all individuals in the pedigree were selected and founder representation and pedigree depth were calculated. Founder representation of breeding individuals was analysed across the three most recent breeding seasons, including 2016-2017 ( $n = 52$ ), 2017-2018 ( $n = 68$ ), and 2018-2019 ( $n = 45$ ). Analyses were performed by breeding season, with some individuals excluded from analyses (Table 2.5).

Table 2. 5: The number of individuals included in founder analyses (*n*), the number of wild and captive pairs for each season, and list of individuals that were excluded from founder analyses.

Breeding Season	<i>n</i>	# Wild Pairs	# Captive Pairs	Individuals Excluded from Analyses
2016-2017	52	22	5	BKW/OY and BKOY/BKR wild fostered chicks, with none surviving
2017-2018	68	32	4	BKWG/BK— and BKWR/O— lost bands
2018-2019	45	23	3	BKWG/BK— lost bands; 3 additional pairs of unknown parents

Assuming that similar pairs of individuals may breed in the 2019-2020 breeding season, placeholder offspring (one male, and one female) were generated for each known breeding pair in the 2018-2019 breeding season (*n* = 44 offspring). Mean kinship values for these individuals were produced to assess whether certain offspring should be targeted for captive pairing.

### 2.3.4.3: Kakī Pedigree Results

The entirety of the kakī pedigree as of July 2018 includes 2,680 wild and captive individuals recorded from 1977-present, with 94 founders and 12.4 founder genome equivalents (Table 2.6). The oldest recorded kakī is 16.4 years old. The sex ratio (i.e., % males) in the kakī pedigree is skewed, as most individuals in the population are of unknown sex. Gene diversity is high (0.96), and mean inbreeding (0.03) and average mean kinship (0.04) are low (Table 2.6). Only 55% of the pedigree can be traced back to founders, indicating that many unknown individuals exist in the kakī pedigree.

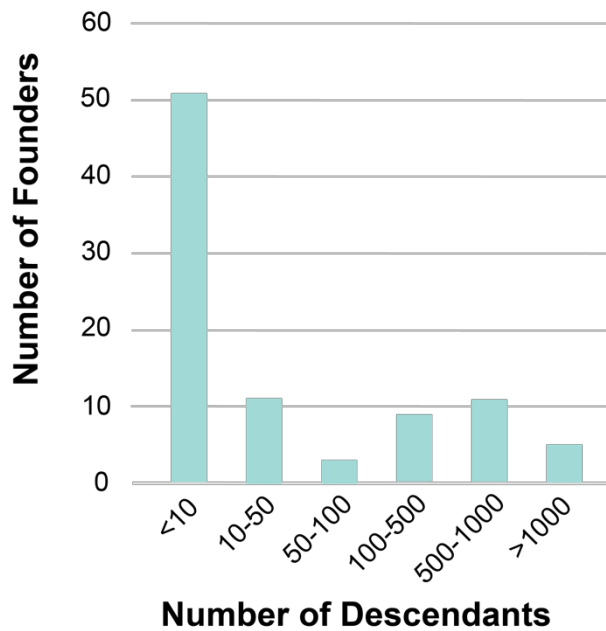
Table 2. 6: Pedigree descriptive statistics for kakī, as produced by PMx, for all known individuals. \*Note, the sex ratio is skewed as most individuals in the population are unknown.

Diversity Statistic	All Individuals
#Individuals	2680
Sex Ratio (% Males)	0.25*
Age Range	<24 years
Gene Diversity	0.96
# Founders	94
Founder Genome Equivalents	12.40
Mean Inbreeding	0.03
Average MK	0.04
Mean Generation Time	4.59
% Ancestry Known	55
% Analytic Known	100
Ne/N	0.10

There are 94 founders in the entire pedigree, including 32 wild (i.e., unbanded) individuals, 44 individuals of unknown or 'mult' parentage (i.e., multiple, usually defined by a trio of potential breeders), 15 of hybrid origins, and 3 of both unknown and wild parentage (Table 2.7). Many of these individuals were 'sourced' (i.e., originally identified and recorded as contributors to the breeding programme of wild or unknown parentage) after the conservation breeding programme initiated in 1981. More than half of these founders have relatively few descendants ( $< 10$ ), while a few founders have many descendants ( $> 100$ , Figure 2.5). Pedigree depth ranges from 0-5.4 generations across the entire pedigree, but is on average 3.39 generations deep for breeding individuals from the 2018-2019 breeding season.

*Table 2. 7: All founders represented throughout the pedigree (Unique ID), along with the year sourced, founder representation (rep.), and number of descendants (Descend.). Founders include individuals of wild parentage (Wild), unknown parentage (UNK), multiple parentage (Mult), and individuals of hybrid parentage, with hybrid node provided in parentheses.*

Parents	Unique ID	Year Sourced	Founder Rep.	Descend.	Parents	Unique ID	Year Sourced	Founder Rep.	Descend.
Wild	4	1983	0.043	806	UNK	936	1984	0.0231	802
Wild	18	1979	0.1048	1644	UNK	939	1990	0.0002	1
Wild	21	1979	0.0002	1	UNK	941	1990	0.0033	18
Wild	23	1986	0.0342	521	Wild & UNK	942	1988	0.0019	9
Wild	27	1979	0.0295	891	UNK	945	1990	0.001	5
UNK	36	1981	0.0002	1	UNK	954	1988	0.0023	11
Wild	196	1982	0.0025	12	Mult	976	1985	0.1004	1515
Wild	207	1986	0.0002	1	Wild	978	1990	0.0373	504
UNK	209	1989	0.0017	11	UNK	991	1981	0.0002	1
Wild & UNK	213	1989	0.0209	208	Wild	1069	1986	0.0002	1
UNK	217	1989	0.0475	1105	UNK	1082	1991	0.0012	6
UNK	224	1984	0.0068	384	UNK	1091	1991	0.0051	29
UNK	273	1996	0.003	18	UNK	1135	1981	0.0008	4
UNK	341	1989	0.026	533	UNK	1153	1989	0.0455	786
Wild & UNK	348	1989	0.0008	4	UNK	1297	2004	0.008	98
Wild	366	1985	0.0189	487	UNK	1600	2012	0.0138	185
Wild	373	1980	0.0016	8	UNK	1907	2012	0.0002	1
UNK	379	1987	0.0171	254	Poaka (A)	Mate Of 0015	1993	0.0004	2
UNK	384	1981	0.004	83	Mult	Mate Of 0143	2006	0.0025	46
Wild	391	1988	0.039	807	Wild	Mate Of 0196	1985	0.0006	3
UNK	396	1996	0.0008	4	Wild	Mate Of 0217	1989	0.0229	779
Wild	398	1989	0.0753	1368	Wild	Mate Of 0224	1987	0.0004	2
Wild	436	1983	0.0064	385	Hybrid (G)	Mate Of 0273	1998	0.0004	2
UNK	438	1986	0.0046	23	Hybrid (I)	Mate Of 0292	2000	0.0008	4
Wild	441	1983	0.0021	10	Wild	Mate Of 0341	1994	0.026	533
Wild	446	1982	0.0002	1	Wild	Mate Of 0348	1993	0.0008	4
Wild	566	1979	0.0027	87	Wild	Mate Of 0366	1993	0.0189	487
UNK	571	1987	0.0231	802	Wild	Mate Of 0436	1986	0.0016	13
UNK	604	1994	0.0994	1510	Wild	Mate Of 0438	1989	0.0005	3
UNK	605	1994	0.0006	3	Hybrid (H)	Mate Of 0441	1989	0.0008	4
UNK	613	1986	0.0008	4	Poaka (A)	Mate Of 0446	1981	0.0002	1
UNK	620	1994	0.0021	12	Wild	Mate Of 0571	1989	0.0002	1
UNK	621	1994	0.0088	371	Hybrid (C)	Mate Of 0657	1981	0.0008	4
UNK	622	1981	0.0008	4	Wild	Mate Of 0690	1989	0.0005	3
UNK	623	1993	0.0004	2	Hybrid (F)	Mate Of 0735	1999	0.0002	1
UNK	657	1984	0.0008	4	Hybrid (F)	Mate Of 0933	1994	0.0017	10
UNK	690	1987	0.0005	3	Hybrid (I)	Mate Of 0934	1993	0.0002	1
UNK	742	1986	0.0006	3	Hybrid (G)	Mate Of 0935	1992	0.0008	4
UNK	748	1987	0.0004	2	Hybrid (H)	Mate Of 0939	1992	0.0002	1
UNK	787	2005	0.0019	9	Wild	Mate Of 0945	1993	0.001	5
Wild	814	1985	0.0024	20	Hybrid (H)	Mate Of 0978	1992	0.0011	7
Wild	815	1988	0.0021	10	Wild	Mate Of 0991	1989	0.0002	1
UNK	909	1992	0.0008	4	Poaka (A)	Mate Of 1081	1981	0.0002	1
UNK	915	1985	0.0015	8	Hybrid (I)	Mate Of 1153	1993	0.0004	2
UNK	933	1990	0.0017	10	UNK	Mate Of 1377	2012	0.0008	4
UNK	934	1990	0.0255	486	Wild	Mate Of 1847	2014	0.0004	2
UNK	935	1990	0.0008	4	Hybrid (?)	Mate Of 2113	2017	0.0002	1



*Figure 2. 5: Number of founders that contributed to descendants across the entire pedigree, with the number of descendants binned into groups.*

There have been 29 founders represented across the last three breeding seasons in the wild and captive populations (Figure 2.6). Of these founders, 12 are descended from unbanded or wild parents, while 16 have unknown or ‘mult’ (i.e., trio) parents and one has both unknown and wild parents. Many of these founding lineages ( $n = 22$ ) have been represented in captivity over the past three seasons by 3-5 pairs. Some of these founders have very high representation (e.g., Unique ID: 0018, 0398, 0604, 0976) while others have very low representation (e.g., Unique ID: 0379, 0384, 0566, Mate of 0145; Figure 2.6). Two rare lineages (Unique ID 0384 and 0566) were captured in the 2016-2017 and 2017-2018 seasons by three breeders: Unique ID: 1810, 1811, and 2008. These three individuals did not breed in the 2018-2019 breeding season, and therefore founding lineages 0384 and 0566 were not represented in this season.

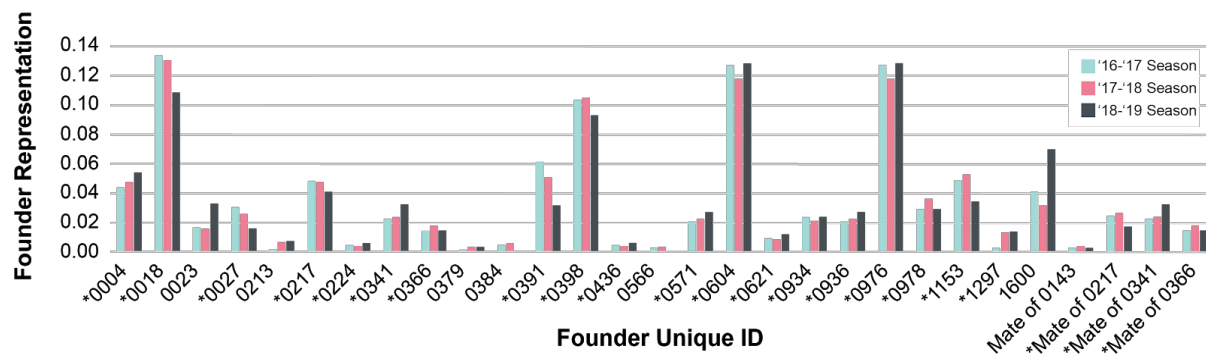


Figure 2. 6: Founder representation across all breeding individuals over the past three breeding seasons. Founders with an asterisk (\*) by their ID have been captured in the captive population over the past three seasons.

Amongst offspring from the 2018-2019 breeding seasons, all individuals had non-zero mean kinship values, indicating shared ancestry in the population. Offspring of pairs 1343 & 1851, 1856 & 2005, and 1461 & 0633 had the lowest mean kinship values (Table 2.8). These offspring are generally represented by many founders, including founders that are under-represented (e.g., Unique ID: 0224, 0379, 0436) and over-represented (e.g., Unique ID: 0018, 0398, 0976) in recent seasons.

Table 2. 8: Mean kinship for offspring of breeding pairs from the 2018-2019 breeding season.

Sire		Dam		Offspring
Bands	Unique ID	Bands	Unique ID	MK
BK(W)/OY	1343	BKOG/BKR	1851	0.0519
BKOG/GO	1856	BKYBK/WG	2005	0.059
WBK/GO	1461	(RO)/GG	633	0.0655
BKRW/GG	2475	BKYR/BKO	2086	0.0713
BKBKR/YO	1729	BKOY/YW	1980	0.0716
BKYO/YY	2083	BKYY/GBK	2126	0.073
BKBK/YW	1334	YG/RBK	1042	0.0743
(R)R/WR	687	BKWY/RR	2284	0.0752
BKOG/BKBK	1848	GBK/WR	260	0.0786
BKRY/WO	2529	BKRO/BKR	2435	0.0788
BKOBK/OO	1826	BKOY/RW	1968	0.079
BKYBK/WR	2007	BKWBK/WBK	2174	0.0799
BKYY/OR	2135	(BK)OR/OG	1903	0.0806
BKYO/YG	2079	BKYY/OBK	2132	0.0814
OW/O(Y)	1366	BKWBK/OW	2166	0.0816
BKYR/GR	2093	BKWO/BKR	2225	0.0827
BKBKY/WG	1792	BKYO/GW	2058	0.0831
BKWO/YG	2389	BKRBK/WO	2329	0.0849
BKYY/GG	2127	BKYY/RY	2141	0.0873
GBK/WG	258	BKBKY/WBK	1791	0.0924
BKWO/WO	2384	BKYG/GBK	2024	0.0928
BKRBK/RBK	2321	BKYR/YO	2116	0.0984

#### 2.3.4.4: Kakī Pedigree Discussion

This report has produced the first functional multigenerational pedigree for kakī, including individuals from the late 1970's to the present. Founder analyses show that of the 94 original founders in the population, only 29 are represented in the most recent breeding seasons. However, many of these original founders were hybrids that have been selected against over the past 20 years, or kakī with unknown (as opposed to unbanded, or wild) parentage sourced well after the conservation breeding programme initiated in 1981. These kakī of unknown parentage are likely to be offspring of captive-bred or captive-reared individuals that are already known within the population.

The large number of founders with unknown parents can be expected in intensively managed wild populations, as wild parents can be more difficult to track and identify outside the confines of captivity (Haig & Ballou 2002). For example, while wild kakī are identified with colour bands, there are instances where the colour bands are difficult to interpret when bands are sun-bleached, caked in mud, or are lost (Department of Conservation, *personal comm.*). As a result, offspring from these individuals will have at least one unknown parent listed in the pedigree, and subsequent estimates of kinship and inbreeding for descendants of this unknown individual may be downwardly biased (Russello & Amato 2004; Ivy *et al.* 2016).

In recent years, the microsatellite panel developed for kakī (Steeves *et al.* 2008) has been used to confirm parentage between uncertain individuals and a list of candidate parents using an allele-mismatch exclusion analysis (Jones & Ardren, 2003). While it is possible for an allele-mismatch exclusion analysis to provide false positives with null alleles (Dakin & Avise 2004) or false negatives due to common alleles across species with low allelic

diversity (Taylor 2015), this approach remains practical for excluding incorrect parents in recent years (Overbeek *et al.* 2017; Overbeek 2019). To minimise unknown individuals in the pedigree moving forward, we recommend the continued use of genetics for exclusion when parentage is uncertain. For example, during the 2017-2018 and 2018-2019 breeding season, individuals with band combinations BKWG/BK— and BKWR/O— are listed as unknown, as both have dropped bands (denoted with a dash “—”). Given the University of Canterbury maintains a feather bank of all managed kakī, it is possible to test offspring from these nests against all individuals with BKWG/BK and BKWR/O bands, to exclude all possible parents from consideration. However, for individuals without a list of suppositional parents, we expect that precise empirical estimates of pairwise relatedness between all breeding individuals in a population using genome-wide single nucleotide polymorphisms (i.e., SNPs; Goudet *et al.* 2018) will be helpful for management (see **Chapters 4 and 5**).

Beyond unknown individuals, wild founders are also treated as unrelated in pedigree analyses (Ballou 1983). Kakī wild founders are likely to be related to one another, given the population dropped to only 23 birds in the early 1980’s (Maloney & Murray 2001). Previous research indicates that pedigrees of substantial depth ( $\geq$  five generations represented) are more robust regarding this assumption and provide more accurate measures of kinship and inbreeding (Balloux *et al.* 2004; Pemberton *et al.* 2008; Rudnick & Lacy 2008). However, the kakī pedigree is still relatively shallow and susceptible to underestimated kinship and inbreeding due to this founder assumption. While there are no DNA or tissue samples available for most wild founders in the population, we do have tissue samples for many of the breeding individuals over the past 15+ years. We anticipate that empirical estimates of pairwise relatedness between recent breeding individuals can be helpful to the kakī pedigree, as they can provide realised proportions of genome sharing between



individuals without the founder bias (Speed & Balding 2015; **Chapters 5 and 6**). While the panel of microsatellites developed for kakī has been useful for allele mismatch exclusion analyses, we expect that using genome-wide SNPs will provide more precise estimates of pairwise relatedness over microsatellites (Attard *et al.* 2018), especially in genetically depauperate species where microsatellite allelic diversity is low (Taylor 2015; Taylor *et al.* 2015).

In the absence of genomic information for all breeding pairs in the 2018-2019 breeding season, the best evidence we have for prioritising individuals for pairings comes from the pedigree. Analyses suggest that offspring from pairs with low mean kinship can be targeted for the formation of new prospective pairs in the 2019-2020 breeding season, should the same individuals breed next season. There were also two under-represented founders that were not captured in the 2018-2019 breeding season: Unique ID 0384 and 0566. If individuals represented by these founders (i.e., Unique ID: 1810, 1811, or 2008) breed during the 2019-2020 breeding season, we recommend their offspring be prioritised as candidates for pairing. Because pedigree-based kinship values for these individuals may be biased by individuals of unknown parentage coupled with low pedigree depth, we recommend empirical estimates of relatedness between candidates be used to inform pairing next breeding season (see **Chapters 5 and 6**).

### *2.3.5: Concluding Remarks*

This chapter developed pedigree resources for two species with conservation breeding programmes. This is an exciting advancement, as both the Kākāriki Karaka Recovery Group and the Kakī Recovery Group can use these tools inform management of breeding pairs in the future. We expect as these pedigrees continue to develop, they may be used to assess other questions in conservation. For example, it is possible

to use pedigrees to assess whether inbreeding has negative fitness consequences across different life history stages (e.g., hatching success, fledging success, and long-term survivorship), as has been done in other pedigreed species in Aotearoa, including takahē (*Porphyrio hochstetteri*; Grueber *et al.* 2008) and black robin (*Petroica traversi*; Kennedy *et al.* 2013).

While these pedigrees have provided helpful information for the conservation management of both species, the kākārīki karaka and kakī pedigrees can benefit from the incorporation of empirical estimates of pairwise relatedness in regards to individuals of unknown parentage. For kakī, we anticipate that microsatellites may continue to be useful for excluding parentage when a panel of possible parents is available (Jones & Ardern 2002). If unknown individuals with no suppositional parents are incorporated into the pedigree — as is the case with Poldark in the kākārīki karaka breeding programme, and occasional unknown birds in the kakī breeding programme — estimates of empirical kinship (i.e., relatedness) between all breeding individuals within the population can be used to avoid bias from the assumption of unrelated founders. While microsatellite panels are developed for both species (Steeves *et al.* 2008; Andrews *et al.* 2013), genome-wide SNPs will provide more precise estimates of relatedness over microsatellite markers (Taylor 2015; Taylor *et al.* 2015; Attard *et al.* 2018; see **Chapter 5**). Moving forward, we recommend the development of genomic resources in both species and investigation on how to best incorporate these estimates into pedigree-based management (**Chapter 6**).

### 2.3.6: References

Andrews BJ (2013) *Conservation genetics of the orange-fronted kākārīki (Cyanoramphus malherbi)* (Unpublished Master's of Science Thesis). University of Canterbury, Christchurch, New Zealand.

- Andrews BJ, Hale ML, Steeves TE (2013) Characterisation of microsatellite loci in the critically endangered orange-fronted kākāriki (*Cyanoramphus malherbi*) isolated using genomic next-generation sequencing. *Molecular Ecology Resources*, **5**, 235–237.
- Attard CRM, Beheregaray LB, Möller LM (2018) Genotyping-by-sequencing for estimating relatedness in nonmodel organisms: avoiding the trap of precise bias. *Molecular Ecology Resources*, **18(3)**, 381–390.
- Ballou, JD (1983) Calculating inbreeding coefficients from pedigrees. In *Genetics and Conservation*. Benjamin Cummings Publishing Co. 509–516.
- Ballou JD, Lacy RC (1995). Identifying genetically important individuals for management of genetic variation in pedigreed populations. In *Population Management for Survival and Recovery*. Columbia University Press. 78–111.
- Ballou JD, Lees C, Faust LJ, Long S, Lynch C, Bingaman LL, Foose TJ (2010) Demographic and genetic management of captive populations. In *Wild Mammals in Captivity: Principles and Techniques for Zoo Management*. University of Chicago Press. 219–252.
- Balloux F, Amos W, Coulson T (2004) Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, **13**, 3021–3031.
- Bergner LM, Jamieson IG, Robertson BC (2014) Combining genetic data to identify relatedness among founders in a genetically depauperate parrot, the kakapo (*Strigops habroptilus*). *Conservation Genetics*, **15**, 1013–1020.
- BirdLife International (2018) *Cyanoramphus malherbi*, Malherbe’s Parakeet. *The IUCN Red List of Threatened Species*.
- Brekke P, Bennett PM, Santure AW, Ewen JG (2011) High genetic diversity in the remnant island populations of hihi and the genetic consequences of re-introduction. *Molecular Ecology*, **20**, 29–45.
- Crow JF, Kimura M (1970) *An introduction to population genetics theory*. Burgess Publishing Co., Minneapolis, Minnesota.
- Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.
- de Villemereuil P, Rutschmann A, Lee KD, Ewen JG, Brekke P, Santure AW (2019) Little adaptive potential in a threatened passerine bird. *Current Biology*, **29**, 889–894.
- Elliott GP, Merton DV, Jansen PW (2001) Intensive management of a critically endangered species: the kakapo. *Biological Conservation*, **99**, 121–133.

Faust LJ, Bergstrom YM, Thompson SD, Bier L (2018) PopLink version 2.4. Available at: <https://www.lpzoo.org/conservation-science/projects/poplink>.

Forsdick NJ, Cubrinovska I, Massaro M, Hale ML (2017) Genetic diversity and population differentiation within and between island populations of two sympatric *Petroica* robins, the Chatham Island black robin and tomtit. *Conservation Genetics*, **18**, 275-285.

Giglio RM, Ivy JA, Jones LC, Latch EK (2016) Evaluation of alternative management strategies for maintenance of genetic variation in wildlife populations. *Animal Conservation*, **19**, 380-390.

Glaser A, van Klink P, Elliott G, Edge K (2010) Whio/blue duck (*Hymenolaimus malacorhynchos*) recovery plan. Department of Conservation, Wellington: New Zealand.

Goudet J, Kay T, Weir BS (2018) How to estimate kinship. *Molecular Ecology*, **27**, 4121-4135.

Grueber CE, Jamieson IG (2008) Quantifying and managing the loss of genetic variation in a free-ranging population of takahe through the use of pedigrees. *Conservation Genetics*, **9**, 645-651.

Hagen EN, Hale ML, Maloney RF, Steeves TE (2011) Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the Black Stilt *Himantopus novaezelandiae*. *Ibis*, **153**, 556-561.

Haig SM, Ballou JD (2002) Pedigree analyses in wild populations. In *Population Viability Analyses*. University of Chicago Press. 388-405.

Hammerly SC, de la Cerda DA, Bailey H, Johnson JA (2016) A pedigree gone bad: increased offspring survival after using DNA-based relatedness to minimize inbreeding in a captive population. *Animal Conservation*, **19(3)**, 296-303.

Hogg CJ, Wright B, Morris KM, Lee AV, Ivy JA, Grueber CE, Belov K (2018) Founder relationships and conservation management: empirical kinships reveal the effect on breeding programmes when founders are assumed to be unrelated. *Animal Conservation*, **22(4)**, 348-361.

Ivy JA, Lacy RC (2012) A comparison of strategies for selecting breeding pairs to maximise genetic diversity retention in managed populations. *Journal of Heredity*, **103**, 186-196.

Ivy JA, Putnam AS, Navarro AY, Gurr J, Ryder OA (2016) Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, **107(5)**, 403-412.

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12(10)**, 2511-2523.

- Jones OR, Wang J (2010) Molecular marker-based pedigrees for animal conservation biologists. *Animal Conservation*, **13**, 26-34.
- Kearvell JC, Grant AD, Boon WM (2003) The orange-fronted parakeet (*Cyanoramphus malherbi*) is a distinct species: a review of recent research into its taxonomy and systematic relationship within the genus *Cyanoramphus*. *Notornis*, **50**, 27-35.
- Kennedy ES, Grueber CE, Duncan RP, Jamieson IG (2013) Severe inbreeding depression and no evidence of purging in an extremely inbred wild species — the Chatham Island black robin. *Evolution*, **68**, 987-995.
- Lacy RC (1995) Clarification of genetics terms and their use in the management of captive populations. *Zoo Biology*, **14**, 565-577.
- Lacy RC (2009) Stopping evolution: genetic management of captive populations. In *Conservation genetics in the age of genomics*. Amato G, DeSalle R, Ryder OA, Rosenbaum HC (Eds). New York: Columbia University Press, 58-81
- Lacy RC, Ballou JD, Pollak JP (2012) PMx: software package for demographic and genetic analysis and management of pedigreed populations. *Methods in Ecology and Evolution*, **3**, 433-437.
- Maloney RF, Murray DP (2001) *Kakī Recovery Plan 2001–2011*. Department of Conservation, Wellington, New Zealand.
- Miskelly CM, Powlesland RG (2013) Conservation translocations of New Zealand birds, 1863-2012. *Notornis*, **60**, 3-28.
- Orr-Walker T, Adams KJ, Roberts LG (2015) *A strategic plan for kea conservation*. Kea Conservation Trust, Queenstown, New Zealand.
- Ortiz-Catedral L, Brunton DH (2009) Notes on the diet of the critically endangered orange-fronted parakeet (*Cyanoramphus malherbi*) on Maud Island. *New Zealand Journal of Zoology*, **36**, 385-388.
- Overbeek AL, Hauber ME, Brown E, Cleland S, Maloney RF, Steeves TE (2017) Evidence for brood parasitism in a critically endangered Charadriiform with implications for conservation. *Journal of Ornithology*, **158**(1), 333-337.
- Overbeek AL (2019) *Evidence for brood parasitism in a critically endangered Charadriiform with implications for conservation* (Unpublished Honours Thesis). Stanford University, Stanford, United States of America.
- Pemberton JM (2004) Measuring inbreeding depression in the wild: the old ways are the best. *Trends in Ecology and Evolution*, **19**(12), 613-615.

- Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society B*, **275**, 613-621.
- Pierce RJ (1984) Plumage, morphology, and hybridisation of New Zealand stilts *Himantopus* spp. *Notornis*, **31**, 106-130.
- Putnam AS, Ivy JA (2014) Kinship-based management strategies for conservation breeding programs when pedigrees are unknown or uncertain. *Journal of Heredity*, **105**, 303-311.
- Rudnick JA, Lacy RC (2008) The impact of assumptions about founder relationships on the effectiveness of conservation breeding strategies. *Conservation Genetics*, **9**, 1439-1450.
- Russello MA, Amato G (2004) Ex situ population management in the absence of pedigree information. *Molecular Ecology*, **13**, 2829-2840.
- Sanders MD, Maloney RF (2002) Causes of mortality at nests of ground-nesting birds in the Upper Waitaki Basin, South Island, New Zealand: a 5-year video study. *Biological Conservation*, **106**, 225-236.
- Smith DHV, Converse SJ, Gibson KW, Moehrensclager A, Link WA, Olsen GH, Maguire K (2011) Decision analysis for conservation breeding: maximizing production for reintroduction of whooping cranes. *The Journal of Wildlife Management*, **75(3)**, 501-508.
- Species360 (2017) Single Population Analysis and Record Keeping System (SPARKS). Available at: <http://www.vortex10.org/sparks16.html>
- Species360 (2019) Zoological Information Management System (ZIMS). Available at: [zims.Species360.org](https://zims.Species360.org).
- Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? *Nature*, **16**, 33-44.
- Steeves TE, Hale ML, Gemmell NJ (2008) Development of polymorphic microsatellite markers for the New Zealand black stilt (*Himantopus novaezelandiae*)\_ and cross-amplification in the pied stilt (*Himantopus himantopus leucocephalus*). *Molecular Ecology Resources*, **8**, 1105-1107.
- Steeves TE, Maloney RF, Hale ML, Tylianakis JM, Gemmell NJ (2010) Genetic analyses reveal hybridization but no hybrid swarm in one of the world's rarest birds. *Molecular Ecology*, **19**, 5090-5100.
- Taylor HR (2015) The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and Evolution*, **5(15)**, 3140-3150.

- Taylor HR, Kardos MD, Ramstad KM, Allendorf FW (2015) Valid estimates of individual inbreeding coefficients from marker-based pedigrees are not feasible in wild populations with low allelic diversity. *Conservation Genetics*, **16**, 901-913.
- van Heezik Y, Lei P, Maloney R, Sancha E (2005) Captive breeding for reintroduction: influence of management practices and biological factors on survival of captive kaki (black stilt). *Zoo Biology*, **24**, 459-474.
- Willoughby JA, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, Dewoody A (2015) The impacts of inbreeding, drift and selection on genetic diversity in conservation breeding populations. *Molecular Ecology*, **24**, 98-110.
- Wright S (1922) Coefficients of inbreeding and relationship. *The American Naturalist*, **56**, 330-338.

## Chapter 3: How to Bridge the Conservation Genomics Gap

### 3.1: Preamble

**Chapter 2** developed pedigree resources for kakī and kākāriki karaka and answered questions of interest for conservation in both species. These analyses also revealed limitations to pedigrees regarding individuals of unknown parentage that are routinely incorporated into conservation breeding programmes. For these individuals, we anticipate that empirical estimates of relatedness using genome-wide SNPs will be the most informative approach to capture realised relatedness amongst all individuals in these programmes. However, at the conception of this thesis (late 2015), relatively few conservation researchers had transitioned from genetic to genomic technologies, despite genomic technologies being available. An article by Shafer *et al.* (2015) coined this paucity the ‘conservation genomics gap’ and identified barriers to genomic research, including poor research-practitioner relationships, a lack of funding and/or genomic resources for non-model threatened species, and a need for bioinformatic upskilling and expertise.

In **Chapter 3**, we suggest an approach for conservation researchers to transition towards genomic technologies by building mutually-beneficial relationships with another applied biological discipline with aligned genomic questions: primary industry research. Through these relationships, conservation researchers have the opportunity to upskill in genomic data management and analysis, while primary industry can gain from incorporating biodiversity expertise into species breeding and management.

This chapter is presented as a published manuscript in the journal *Molecular Ecology*, published in late 2016. Since the publication of this article, these relationships with



primary industry research have led to upskilling in bioinformatic expertise needed to complete this thesis. For example, workshops led by Roger Moraga, formally with New Zealand AgResearch and now with Tea Break Bioinformatics Limited, allowed me to upskill in reduced representation and resequencing SNP discovery (**Chapters 4 & 5**). By attending the New Zealand MapNet meetings (e.g., <http://mapnet2019.nz/>), we have had the exposure to primary industry research pipelines, like KGD (Dodds *et al.* 2015) that have been used to estimate relatedness in this thesis (**Chapters 4 & 5**).

Beyond the work shown here, this approach has been used by our wider research group (the Conservation, Systematics, and Evolutionary Research Team, or ConSERT). For example, it led to co-developed and nationally-funded genomic research on kēkēwai/freshwater crayfish (*Paranephrops zealandicus*), which aims to improve conservation, customary and commercial outcomes for this declining mahinga kai species. Moving forward, we anticipate this approach will continue to lead to mutually beneficial outcomes for both disciplines.

The published PDF of this chapter can be found in **Appendix B**, with the Supplemental Materials for this chapter found in **Appendix C**.

### *3.1.1: Contribution Statement*

The concept for this article was conceived by myself and Tammy Steeves. The research was designed in collaboration with all co-authors listed, including Thomas R. Buckley, Rob Elshire, Marie L. Hale, Michael Knapp, John McCallum, Roger Moraga, Anna W. Santure, and Phillip Wilcox. I led this piece, coordinated all perspectives, compiled the literature search, and provided written drafts and figures. All coauthors contributed to revisions. Tammy Steeves, Rob Elshire, and Phil Wilcox contributed perspective boxes.



**Stephanie J. Galla**

### 3.2: Building strong relationships between conservation genetics and primary industry leads to mutually-beneficial genomic advances

Stephanie J. Galla<sup>\*1</sup>, Thomas R. Buckley<sup>2,7</sup>, Rob Elshire<sup>3</sup>, Marie L. Hale<sup>1</sup>, Michael Knapp<sup>4</sup>, John McCallum<sup>5</sup>, Roger Moraga<sup>6</sup>, Anna W. Santure<sup>7</sup>, Phillip Wilcox<sup>8</sup>, and Tammy E. Steeves<sup>1</sup>

<sup>1</sup>School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch, 8140, New Zealand

<sup>2</sup>Landcare Research, Private Bag 92170, Auckland Mail Centre, Auckland, 1142, New Zealand

<sup>3</sup>The Elshire Group, Ltd., 52 Victoria Avenue, Palmerston North, 4410, New Zealand

<sup>4</sup>Department of Anatomy, University of Otago, P.O. Box 913, Dunedin, 9054, New Zealand

<sup>5</sup>Breeding and Genomics, New Zealand Institute for Plant and Food Research, Private Bag 4704, Christchurch, 8140, New Zealand

<sup>6</sup>AgResearch, Ruakura Research Centre, Bisley Road, Private Bag 3115, Hamilton, 3240, New Zealand

<sup>7</sup>School of Biological Sciences, University of Auckland, Auckland 1010, New Zealand

<sup>8</sup>Department of Mathematics and Statistics, University of Otago, P.O. Box 56, 710 Cumberland Street, Dunedin, 9054, New Zealand

#### 3.2.1: Abstract

Several reviews in the past decade have heralded the benefits of embracing high-throughput sequencing technologies to inform conservation policy and the management of threatened species, but few have offered practical advice on how to expedite the transition from conservation genetics to conservation genomics. Here, we argue that an effective and efficient way to navigate this transition is to capitalize on emerging synergies between conservation genetics and primary industry (e.g., agriculture, fisheries, forestry and horticulture). Here, we demonstrate how building strong relationships between conservation geneticists and primary industry scientists is leading to mutually-beneficial outcomes for both disciplines. Based on our collective experience as collaborative New

Zealand-based scientists, we also provide insight for forging these cross-sector relationships.

**Keywords:** conservation genomics, high-throughput sequencing, next-generation sequencing, interdisciplinary research, applied research

### 3.2.2: Introduction

One does not need to read beyond the pages of *Molecular Ecology* to see how emerging technologies are revolutionizing the way we conduct research in ecology and evolutionary biology (i.e., EEB) and conservation biology. This is exemplified by rapid advances in genomics, where in the span of two decades the field of molecular ecology has grown from using Sanger technologies to sequence single target loci to using high-throughput sequencing (HTS) technologies to affordably sequencing entire draft genomes (Narum *et al.* 2013; Payseur & Rieseberg 2016; Tigano & Friesen 2016). When new technologies become available, there is a tendency for reviews to be published heralding their potential to address new and exciting questions. Beyond the value of these reviews, an even more important conversation needs to take place in the peer-reviewed literature: how do we efficiently incorporate new technologies into our research repertoire to make accelerated gains in applied and fundamental science?

The field of conservation genetics is currently in transition given rapid advancements in HTS technologies. Many reviews have highlighted the promise of embracing HTS technologies in conservation (Luikart *et al.* 2003; Kohn *et al.* 2006; Primmer 2009; Allendorf *et al.* 2010; Avise 2010; Frankham 2010a; Ouburg *et al.* 2010; Angeloni 2011; Ekblom & Galindo 2011; Funk 2012; McCormack *et al.* 2013; Narum *et al.* 2013; Steiner *et al.* 2013; Ellegren 2014; McMahon *et al.* 2014; Shafer *et al.* 2015; Andrews *et al.* 2016; Benestan *et al.* 2016; Grueber 2016). However, as recently discussed by Shafer *et al.* (2015, 2016) and

Garner *et al.* (2016), there are a limited (albeit exponentially increasing) number of published empirical studies that apply HTS data to conservation. We are aware of empirical genomic studies in EEB that are applicable to questions in conservation (e.g., Defaveri *et al.* 2013; Hoffman *et al.* 2014; Knief *et al.* 2015; Bérénos *et al.* 2016; Hess *et al.* 2016; Prince *et al.* 2017) and there are many EEB researchers applying their genomics expertise to improve conservation outcomes for threatened species, including two of our co-authors (MK, AWS). In addition to the EEB sphere, there are conservation geneticists (e.g., our co-authors SJG, TRB, MLH, TES) who are successfully venturing into conservation genomics through collaborations with colleagues in another applied discipline well-versed in genomics: primary industry (a collective term referring to scientists in agriculture, fisheries, forestry and horticulture; such as our co-authors RE, JM, RM, PW). Through building these cross-sector relationships, it has become clear that there is immense potential for conservation geneticists and primary industry scientists to collaborate on applied research that addresses aligned questions using similar genomic approaches. In this perspective, we use our experience as a collaborative group of New Zealand-based scientists to argue that building strong relationships between conservation genetics and primary industry can lead to improved genomic outcomes for both disciplines and offer advice on how to best build meaningful cross-sector relationships.

### *3.2.3: Conservation Genetics and Genomics*

Before discussing mutually-beneficial genomic synergies between conservation genetics and primary industry, we feel it is important to first discuss what conservation genetics is, what can be gained by using a genomic approach and what obstacles may impede geneticists from adopting genomic technologies. Conservation genetics is a

subdiscipline of conservation biology (Soulé 1985) which uses genetic data to inform the management of threatened species in collaboration with conservation practitioners (Frankham 1995; Avise 2008; Frankham 2010b; Haig *et al.* 2016). While there is overlap between the fields of conservation genetics and EEB, we distinguish conservation genetics as an applied subdiscipline with direct implications for the management and of threatened species. Many threatened taxa have experienced significant population declines (i.e., demographic bottlenecks, see Keller *et al.* 1994), leading to small populations that are susceptible to genetic factors (i.e., loss of genetic diversity, inbreeding and inbreeding depression) associated with extinction risk (Frankham 1995). Conservation geneticists have traditionally used few targeted neutral genetic markers including mitochondrial sequences, microsatellites and amplified fragment length polymorphisms (AFLPs) to measure inbreeding, relatedness and genetic diversity within threatened populations, estimate population genetic structure and gene flow among threatened populations, delineate species boundaries in threatened taxa and detect hybridisation and introgression between threatened and non-threatened species (Allendorf *et al.* 2010; Ouborg *et al.* 2010).

Advancements in HTS technologies are enabling the development of genomic resources for threatened species including the *de novo* assembly and annotation of high-quality reference genomes (e.g., Li *et al.* 2014, Zhang *et al.* 2014) and characterization of a large number genome-wide markers such as single-nucleotide polymorphisms (SNPs) (e.g., Benestan *et al.* 2015; Kraus *et al.* 2015; Lemay *et al.* 2015). For conservation geneticists who have traditionally used small panels of neutral genetic markers to estimate population genetic parameters above and below the species level, HTS technologies are appealing as they enable an affordable means to discover and genotype a large quantity of genome-wide SNPs (Avise 2010; McCormack *et al.* 2013; Shafer *et al.* 2015) and these large SNP datasets

are more representative of genome-wide variation and can result in higher resolution estimates of population genetic parameters (Väli *et al.* 2008; Ljungqvist *et al.* 2010; Santure *et al.* 2010; Taylor *et al.* 2015). In the field of conservation genetics and EEB, a small but rapidly growing number of empirical studies have demonstrated the utility of genomic markers in estimating population genetic structure and gene flow (Bowden *et al.* 2012; Dierickx *et al.* 2015; Lew *et al.* 2015; Oyler-McCance 2015), estimating relatedness (Béréños *et al.* 2016), measuring genome-wide diversity (Robinson *et al.* 2016) and detecting hybridisation and introgression (Hohenlohe *et al.* 2013). We anticipate even more conservation geneticists will begin to embrace HTS technologies as empirical evidence demonstrating the superiority of using genomic markers to inform conservation decisions grows and the costs of doing so diminishes (Box 1).

**Box 1. The costs of using a conservation genomic approach. *Perspectives are those of Tammy Steeves.***

Since I arrived in New Zealand from Canada in 2004, I have had the privilege of developing conservation genetic management recommendations in collaboration with several Department of Conservation recovery or specialist groups to assist the recovery of endemic taonga (treasured) bird species. To date, these recommendations have been predominantly based on genetic markers, namely mitochondrial sequences or microsatellite genotypes (e.g., Steeves *et al.* 2010; Hagen *et al.* 2011; Overbeek *et al.* *In press*). In collaboration with primary industry colleagues in the MapNet community (see Box 2 & 3), I recently assessed the direct and indirect costs associated with shifting from a conservation genetic to a conservation genomic approach and decided to develop genomic markers (SNPs) for the endangered tuturuatu/shore plover (*Thinornis novaeseelandiae*; Fig. A) and the critically endangered kakī/black stilt (*Himantopus novaeseelandiae*; Fig. B).

**Tuturuatu/Shore plover** - I was recently invited to be an expert advisor to the Shore Plover Specialist Group. The Specialist Group was interested in sampling captive and wild birds to estimate the extent of population genetic structure, and compare levels of genetic diversity, between captive and wild shore plover. To achieve this, I knew the cost to develop, screen and genotype ~20 polymorphic species-specific microsatellites for 94 individuals (~10K NZD) would be more than using a reduced-representation approach to simultaneously discover and genotype >20,000 SNPs for the same number of individuals (Elshire *et al.* 2011; ~8.5K NZD). I also knew it would be possible to expedite the characterisation of SNPs if I was able to use a reference-guided approach. As a member of the Avian Genome Consortium, I was aware bird genomes are small, compact and highly conserved (Zhang *et al.* 2014), and that one of the newly available high quality bird genomes (killdeer, *Charadrius vociferus*) would likely be an appropriate proxy-reference genome for SNP discovery and genotyping in shore plover because both species are members of the Family Charadriidae (Card *et al.* 2014). Thus, the main driver of my decision to embrace a conservation genomic approach for shore plover was to ensure that I could develop a comprehensive postgraduate research project that could deliver pertinent results to the Shore Plover Specialist Group in a timely fashion.



Fig. A. Tuturuatu/Shore plover

**Kakī/black stilt** - As a member of the Kakī Recovery Group, I have used species-specific genetic markers to inform the conservation genetic management of captive and wild kakī populations for many years. For example, I routinely use genetic-based measures of relatedness based on microsatellites to inform captive pairing decisions (as per Hagen *et al.* 2011). However, emerging evidence indicates genetic-based measures are relatively poor indicators of genome-wide diversity, particularly in genetically impoverished species like kakī, and a better indication of genome-wide diversity should be obtained from genomic-based measures of relatedness based on genome-wide SNPs (Taylor *et al.* 2015; Willoughby *et al.* 2015). Thus, the main driver of my decision to generate SNPs for kakī was to establish the Kakī Recovery Programme as an exemplar of 'best practice' conservation genomic management.



Fig. B. Kakī/Black stilt

The paradigm underlying many conservation genetic studies is that a genetically diverse population as measured by neutral genetic markers is also likely to be functionally diverse (Bataillon *et al.* 1996) and therefore better able to adapt to environmental change (Frankham 2005). While many have aspired to move past this paradigm, it remains entrenched in most conservation genetic studies that use neutral markers (Caballero & García-Dorado 2013; Vilas *et al.* 2015). As a result of the lack of empirical data on functional genetic diversity in species of conservation interest, beyond studies that include

immunocompetence genes like those in the major histocompatibility complex and toll-like receptors (reviewed in Grueber 2016), it has been difficult to assess the validity of this conservation genetic paradigm. Further, even if supported by empirical data, neutral genetic data might not be a suitable proxy for functional genetic data for threatened species. For example, the translocation of individuals from a large genetically diverse population to supplement a small genetically depauperate population might introduce new genetic diversity (Weeks *et al.* 2011; IUCN 2013), but it might also inadvertently lead to outbreeding depression if source and recipient populations are each locally adapted (Edmands 2007; Frankham *et al.* 2011, but see Frankham 2015; Whiteley *et al.* 2015; He *et al.* 2016).

There is exceptional interest in using a conservation genomics approach to detect regions of the genome that underlie phenotypic variation linked to fitness in threatened populations (i.e., adaptive variation; Luikart *et al.* 2003; Kohn *et al.* 2006; Ouburg *et al.* 2010; Angeloni *et al.* 2011; Harrison *et al.* 2014; Shafer *et al.* 2015). There are several methods available to study adaptive variation, including gene mapping approaches (i.e., genome-wide association studies or GWAS, and quantitative trait loci mapping or QTL; Slate *et al.* 2010; Stapley *et al.* 2010), outlier locus analysis (Luikart *et al.* 2003), and selective sweep mapping (Pardo-Diaz *et al.* 2015). However, determining the genetic basis of phenotypic traits, especially those linked to fitness, is complex, owing to the fact that most fitness-related traits are likely to be controlled by multiple loci (Savolainen *et al.* 2013) and many are likely to be under at least some environmental influence (Falconer & Mackay 1996; Lynch & Walsh 1998). In addition, the success of these approaches is often contingent on large sample sizes (e.g., Ball 2005) which will be challenging to generate for most species of conservation concern.



While there are challenges associated with the detection of adaptive variation in threatened populations (reviewed in Shafer *et al.* 2015), there is potential to answer new questions previously not tractable by employing small sets of targeted genetic markers. In particular, an understanding of the genetic basis of fitness traits will allow more robust predictions of the evolutionary potential of threatened species (Ouberg *et al.* 2010), including a better understanding of genetic trade-offs between traits that might constrain adaptation (Slate *et al.* 2010). Further, identifying loci underlying local adaptation is likely to help identify candidate populations for conservation translocations (Seddon 2010; He *et al.* 2016). Finally, identification of genes responsible for detrimental traits associated with inbreeding depression will have immediate impact on the management of threatened species, especially where matings between individuals are managed (e.g., captive populations; Angeloni *et al.* 2011; Harrison *et al.* 2014; Shafer *et al.* 2015).

Despite having been available for over a decade (Margulies *et al.* 2005), a limited number of publications have applied HTS technologies to conservation (Shafer *et al.* 2015, 2016a; but see Garner *et al.* 2016), with the term ‘conservation genomics gap’ first being used in 2015 to describe the paucity of conservation geneticists using HTS technologies to inform conservation management (Shafer *et al.* 2015). While there are a growing number of examples that show how genomic data is being used to inform conservation decisions (Garner *et al.* 2016; but see Shafer *et al.* 2016; see Figure Appendix C1) and many conservation geneticists who are currently producing HTS datasets, there has been a substantial time lag between when these techniques have become available and uptake by the conservation research community, especially in comparison to other applied genetic disciplines like primary industry (e.g., agriculture, fisheries, forestry, and horticulture; see Figure 3.1). In addition, much of the uptake in conservation biology has been restricted to

threatened wild fish stocks (Garner *et al.* 2016; Shafer *et al.* 2016). Of the 51 articles in Figure 3. 1 classified as ‘conservation genomics’, 30% pertained to the management of declining, over-fished or threatened commercially fished species (e.g., Atlantic salmon, *Salmo salar*; orange-roughy, *Hoplostethus atlanticus*; delta smelt, *Hypomesus transpacificus*), which provides an excellent example of how conservation genomic research can also be relevant to other scientific disciplines including primary industry (e.g., these articles were classified as both ‘conservation genomics’ and ‘primary industry’ in Figure 3.1).

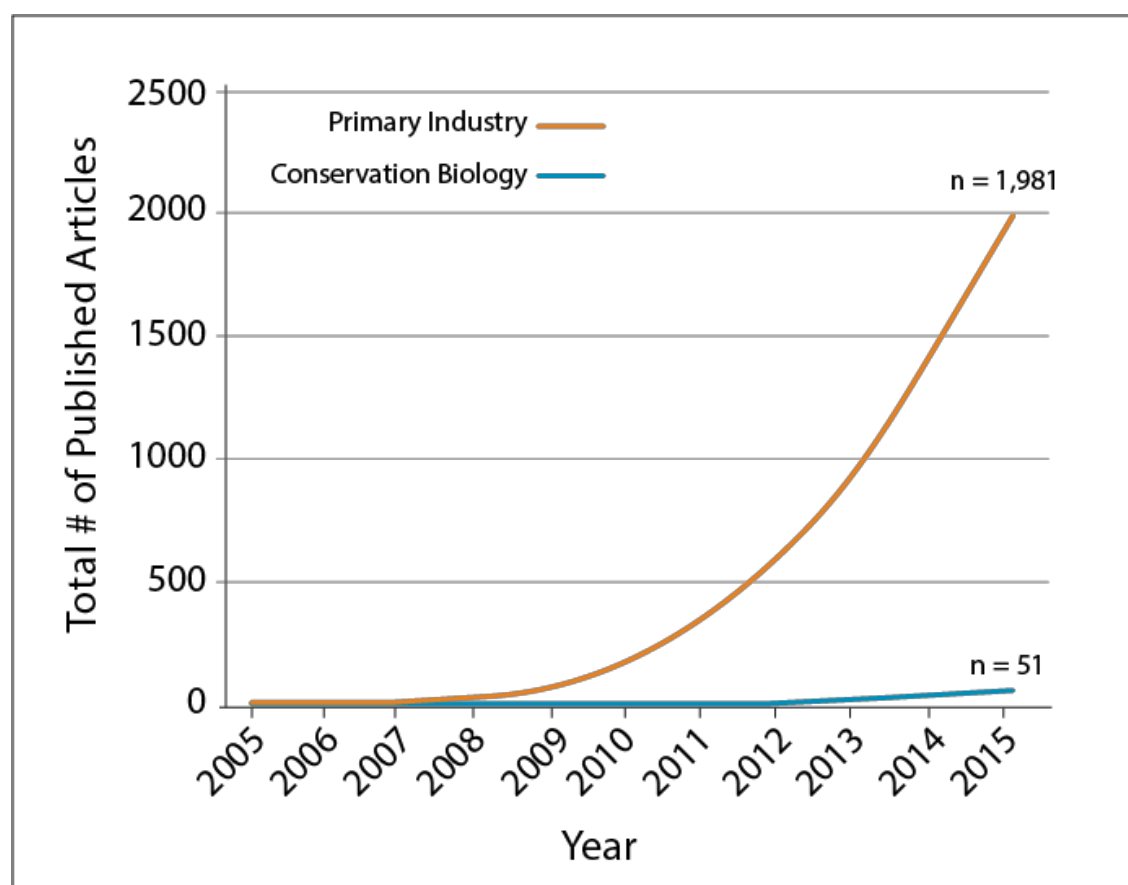


Figure 3. 1: Number of publications using high-throughput sequencing technologies to generate genomic data in conservation (blue line) and primary industry (red line) from 2005-2015. Values for this graph were derived from an ISI Web of Science literature search, using inclusive terminology (see Appendix C for details). Curved lines have been smoothed for ease of interpretation.

Shafer *et al.* (2015) predominantly attribute the conservation genomics gap to a persistent disconnect between academia and real-world conservation issues. We agree strong relationships between academics and conservation practitioners are crucial, but

argue the conservation genomics gap as defined by Shafer *et al.* (2015) is more akin to a 'research-implementation gap' (Knight *et al.* 2008; Hogg *et al.* 2016). Indeed, if strong relationships between academics and conservation practitioners are absent, the likelihood that *any* research will be translated into conservation action is exceptionally low (Haig *et al.* 2016). Here, we predominantly attribute the apparent shortage of conservation geneticists using HTS technologies (i.e., the conservation genomics gap *sensu stricto*) to several interconnected challenges associated with the generation, analysis and interpretation of genomic data.

Prior to identifying these interconnected challenges, we recognise some questions in conservation are still being readily addressed with genetic data (e.g., Dowling *et al.* 2015; Li *et al.* 2015a; Pacioni *et al.* 2015; Trask *et al.* 2015; Cubrinovska *et al.* 2016; Hammerly *et al.* 2016; Overbeek *et al.* 2017). We anticipate studies such as these to persist, at least in the short-term, because existing panels of genetic markers will remain a sufficient low-cost option in some situations (Angeloni *et al.* 2011; McCormack *et al.* 2011; McMahon *et al.* 2014). Although we acknowledge that direct cost can be a factor contributing to the conservation genomics gap, we do not think it underpins it, especially when reduced-representation approaches (e.g., restriction-site associated DNA sequencing, genotyping-by-sequencing, exome capture, and RAD Capture; Baird *et al.* 2008; Elshire *et al.* 2011; Jones & Good 2015; Ali *et al.* 2016) make it possible to characterize tens-of thousands of SNPs in hundreds of individuals for non-model species at a lower cost than developing and screening relatively few novel microsatellite markers (Narum *et al.* 2013; Andrews *et al.* 2016; Box 1). Beyond direct cost, the shortage of high-quality reference genomes is an often cited impediment to SNP discovery and genotyping for non-model species (e.g., Allendorf *et al.* 2010; Ouberg *et al.* 2010; Shafer *et al.* 2015), particularly when approximate SNP

location is of interest (e.g., Kardos *et al.* 2015). However, an ever increasing number of high-quality and high-coverage genomes are becoming available (Ellegren 2014). It has also become apparent that low-coverage draft genomes (sometimes referred to as ‘landing-pad’ or ‘skim’ genomes), or even highly-quality and high-coverage genomes of closely related taxa, can enable reference-guided mapping assembly and SNP characterization in some taxa (Card *et al.* 2014; Wang *et al.* 2014). The lack of bioinformatic expertise and pipelines required to analyze large population genomic datasets has also been frequently cited as a challenge that precludes the use of HTS technologies in conservation (e.g., McCormack *et al.* 2013; Shafer *et al.* 2015). Steep analytical learning curves are generally associated with new technologies, particularly for rapidly advancing fields like genomics where bioinformatic expertise is needed to analyse large genomic datasets. However, the analysis of large population genomic datasets is no longer exceptional. For example, in regards to SNP discovery and genotyping alone, several comprehensive bioinformatic pipelines are readily available (e.g., Puritz *et al.* 2014; Glaubitz *et al.* 2014; Herten *et al.* 2015; Sovic *et al.* 2015; Melo *et al.* 2016).

Depending on the conservation genetics project at hand, one or a combination of the challenges listed above might impede conservation geneticists from transitioning to HTS technologies. Given the recent developments in HTS technologies and the potential it has for benefitting conservation outcomes, we suggest it is time for researchers to start sharing practical advice on how to expedite the transition from conservation genetics to conservation genomics. Here, we argue that an effective and efficient way to navigate the conservation genomics gap is to capitalize on emerging synergies between conservation genetics and primary industry, and demonstrate how building strong relationships between these two disciplines is leading to mutually-beneficial genomic outcomes.

### 3.2.4: Strong Relationships Lead to Mutually Beneficial Genomic Advances

Conservation geneticists are skilled at building strong relationships in an interdisciplinary landscape to improve conservation outcomes (Haig *et al.* 2016; Hogg *et al.* 2016). However, by pushing the boundaries of the conservation 'silo', conservation geneticists will be better able to navigate the conservation genomics gap if they forge novel relationships with scientists that have shared genomic goals, albeit in a different discipline such as primary industry (Figure 3.2). As a discipline, primary industry represents a diverse group of scientists from universities, private institutions and government organisations that apply scientific data to the benefit of primary production output (e.g., meat, fish, eggs, dairy, fruits, vegetables, fibers and timber). Some of the early draft genomes were published to improve commercial outcomes, including rice (Goff *et al.* 2002), red jungle fowl (Hillier *et al.* 2004), silkworm (Xia 2004) and cattle (Schibler *et al.* 2004). With these early reference genomes and the accumulation of massive SNP datasets coupled with phenotypic data, many primary industry scientists have years of expertise with the application of genomic data. Approximately 1,981 HTS studies using genomic data have been published in primary industry from 2005-2015, which outnumbers those produced in conservation biology by more than an order of magnitude (Figure 3.1).

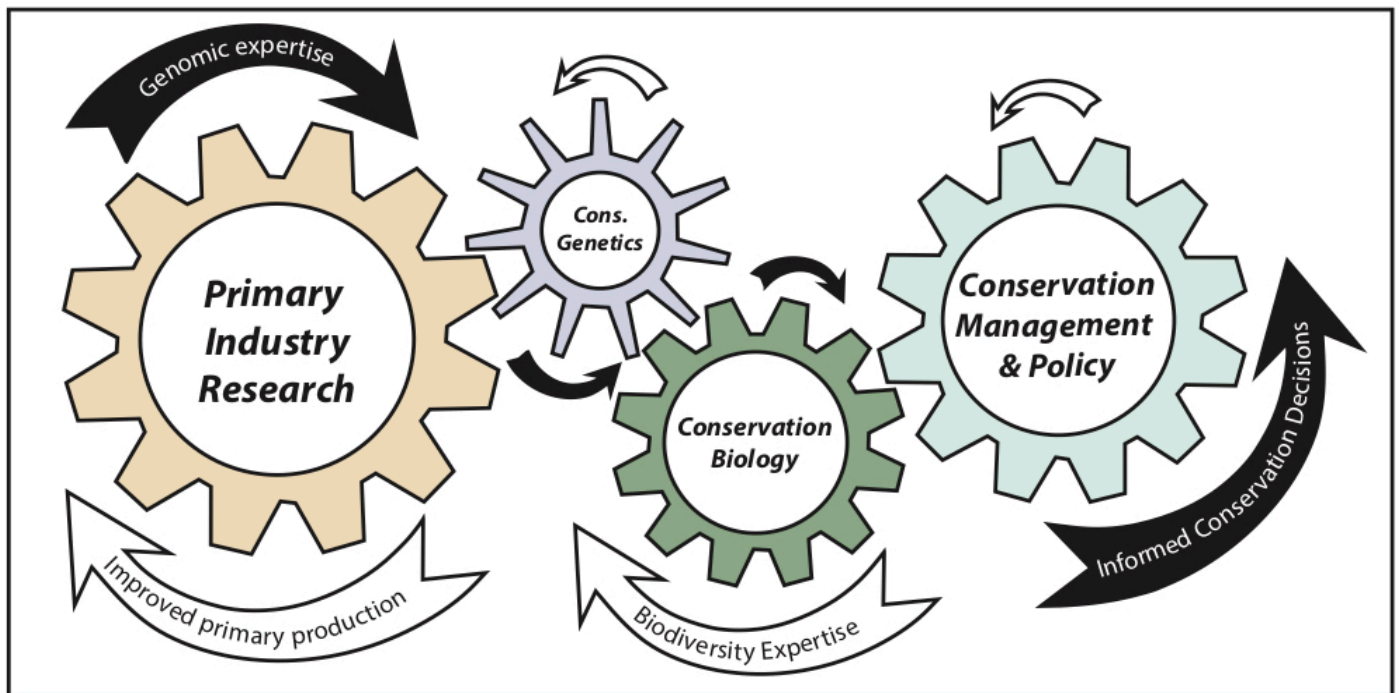


Figure 3. 2 : Simplified schematic detailing how relationships between conservation genetics and primary industry are leading to mutually beneficial outcomes. In black arrows, genomic expertise from primary industry advances conservation genetics, which in turn informs conservation biology and conservation management and policy. In white arrows, biodiversity expertise informs primary industry research, which in turn improves primary production.

Conservation has already benefited from genomic resources provided by primary industry. For example, genomic resources developed for cattle including the draft genome (Schibler *et al.* 2004) and the Bovine SNP chip (Gunderson *et al.* 2005; Steemers *et al.* 2006; Matukumalli *et al.* 2009) have been used to estimate the extent of introgression from cattle to American bison (Halbert *et al.* 2005), measure genomic variation in American and European bison (Pertoldi *et al.* 2009) and develop genomic resources for scimitar-horned and Arabian oryx (Ogden *et al.* 2012). Similarly, genomic resources developed for domestic sheep have been used to describe genome-wide diversity and assess genetic rescue for bighorn sheep (Poissant *et al.* 2009; Miller *et al.* 2012). Of course, there are species of mutual interest to both conservation and primary industry, including species in the fishery and forestry sectors (e.g., Monterey pine, *Pinus radiata* D.Don; New Zealand tōtara,

*Podocarpus spp.*; chinook salmon, *Oncorhynchus tshawytscha*; orange roughy, *Hoplostethus atlanticus*) and therefore genomic resources produced one discipline can be easily used by the other (Dillon *et al.* 2013; Larson *et al.* 2014; da Silva *et al.* 2015; Marshall *et al.* 2015). We anticipate conservation geneticists may opt to use closely-related commercial or model species to inform adaptive variation studies in threatened species, given that gene-mapping approaches are contingent on large sample size (Ball 2005; see discussion above) and the small census size of threatened populations may be inadequate.

Collaborations between conservation geneticists and primary industry scientists are logical because researchers in these two disciplines are beginning to address similar questions in an applied genetic discipline (see Table 3.1). For example, primary industry scientists have been using neutral genome-wide SNPs to calculate inbreeding coefficients in sheep (Li *et al.* 2011), reconstruct parentage assignments in cattle (Hayes *et al.* 2011) and calculate diversity measures for genetic improvement in poultry (Muir *et al.* 2008; Aslam *et al.* 2012). Pipelines that have been used or developed to address these questions in commercial species are likely to be of interest to conservation geneticists, but are sometimes published in discipline-specific peer-reviewed journals such as the *Journal of Dairy Science* or *Plant Biotechnology Journal* (e.g., Allen *et al.* 2012; Li *et al.* 2015b) or reported in outlets that conservation geneticists might not yet be familiar with like *bioRxiv* (e.g., Dodds *et al.* 2015). Similarly, there are some conservation genomic articles from non-academic sources that are not represented in peer-reviewed literature (Garner 2016). These examples highlight how relationships between conservation genetics and primary industry scientists can enable the dissemination of discipline-specific publications and will allow scientists from both disciplines to learn about recently developed pipelines.

Table 3. 1: Common genomic issues facing conservation genetics and parallel examples addressed by scientists in primary industry.

Topic	Challenge for conservation genomics	Examples of corresponding research from primary industries
Polyploid genomes	Developing effective tools for genome-wide SNP discovery and genotyping for plants, invertebrates and some vertebrates with polyploid genomes	Genome-wide SNP studies on polyploids <sup>1</sup> including wheat <sup>2</sup> , cotton <sup>3</sup> , potato <sup>4</sup> , peanut <sup>5</sup>
Genetic basis of adaptive variants	Discovery of variants underpinning traits of relevance to conservation including adaptive variation	Trait mapping for economically important traits using GWAS and QTL mapping <sup>6,7</sup> in rice <sup>8</sup> , dairy cattle <sup>9</sup> , pig <sup>10</sup> , soybean <sup>11</sup> .
Gene copy number variation	Quantifying genome-wide copy number variation and estimating its contribution to phenotypic variation	Quantifying genome-wide copy number variation and estimating its contribution to economically important traits in apple <sup>12</sup> , pig <sup>13</sup> , wheat <sup>14</sup> .
Inbreeding and Relatedness	Measuring inbreeding ( $f$ ), detecting inbreeding depression, and estimating relatedness ( $r$ ) for small populations to maintain evolutionary potential	Measuring inbreeding ( $f$ ), detecting inbreeding depression, and estimating relatedness ( $r$ ) in sheep, <sup>15</sup> pigs <sup>16,17</sup> and salmon <sup>18</sup> to enhance traits for commercial selection

Clevenger *et al.* 2015<sup>1</sup>; Allen *et al.* 2012<sup>2</sup>; Byers *et al.* 2012<sup>3</sup>; Uitdewilligen *et al.* 2013<sup>4</sup>; Bertioli *et al.* 2014<sup>5</sup>; Collard & Mackill 2008<sup>6</sup>; Hu *et al.* 2012<sup>7</sup>; Begum *et al.* 2015<sup>8</sup>; Li *et al.* 2015b<sup>9</sup>; Zhang *et al.* 2015<sup>10</sup>; Zhou *et al.* 2015<sup>11</sup>; Boocock *et al.* 2015<sup>12</sup>; Wang *et al.* 2015<sup>13</sup>; Wuerschum *et al.* 2015<sup>14</sup>; Li *et al.* 2011<sup>15</sup>; Herrero-Medrano *et al.* 2012<sup>16</sup>; Silió *et al.* 2016<sup>17</sup>; Dodds *et al.* 2015<sup>18</sup>

Understanding the genetic basis of desired commercial traits is also a main focus in primary industry (Womack 2005; Tuberosa & Salvi 2006; Sellner *et al.* 2007; Collard & Mackill 2008; Neale & Kremer 2011; Sonah *et al.* 2011; Hu *et al.* 2013). Primary industry has benefitted from collaboration with researchers in human health to determine the genetic basis of phenotypic traits in complex pedigrees and structured populations using QTL mapping and GWAS (George *et al.* 2000; Aulchenko *et al.* 2007; Price *et al.* 2010). In turn, these gene mapping approaches have been successfully applied to understanding the genetic basis of ecologically relevant traits in many wild populations (Schielzeth & Husby 2014). While there are numerous research groups outside of primary industry exploring adaptive variation (e.g., Rietveld *et al.* 2013; Brachi *et al.* 2015; Chaves *et al.* 2016), we



anticipate that conservation geneticists in particular will benefit from forging relationships with primary industry scientists given that both groups work in an applied discipline with species characterised by small effective population sizes. Additionally, there is potential for conservation geneticists to adopt a genomic selection approach (e.g., Heffner *et al.* 2009; Hayes *et al.* 2009) to generate breeding values to inform the selection of individuals for conservation breeding. Lastly, we recognise that both conservation geneticists and primary industry researchers routinely work with species with complex genomes (Clevenger *et al.* 2015) and therefore researchers from these two disciplines have an opportunity to work together and think of creative bioinformatic solutions for species that present bioinformatic challenges (Box 3). Given these commonalities, synergies between both conservation genetics and primary industry can lead to the development of improved HTS techniques and pipelines to address mutual problems in species of both conservation and commercial interest (Box 2; Box 3; Table 3.1).

**Box 2. Retrospective and prospective of Genotyping-by-Sequencing (GBS). *Perspectives are those of Rob Elshire.***

In 2007, I joined the Buckler Lab at Cornell University and the next-generation sequencing revolution simultaneously. My first task was to develop a new library preparation method for the nascent Illumina sequencing platform. The technology was not nearly as robust as it is today and the reads were very short (i.e., 32bp in length). Our challenge was to sequence the non-repetitive fraction of the maize genome. To do that we used a combination of digestion by restriction enzymes and gel based size selection to exclude the repetitive fraction. The data generated formed the basis for the first Maize Hapmap paper (Gore *et al.* 2009). When that project neared completion, I was tasked with building a low-cost, high-throughput genotyping method as an extension of my previous work. The overall goal was to develop a genotyping system that would allow simultaneous marker discovery and genotyping and also address the issue of marker discovery bias. Other researcher groups at the time were developing similar methods, as there was a high demand for an affordable and reproducible method of genotyping and it was the next logical thing to try. One aim was to provide enough genetic markers at the right price point to enable plant breeding by genomic selection. To maximise the benefit of our work and encourage others to take what we did and create new methods appropriate for new questions, we made our work openly available. The resultant genotyping-by-sequencing (GBS) method was published in *PLoS One* in 2011 (Elshire *et al.* 2011).

We achieved our goal of developing a new genotyping method that was inexpensive, both in terms of cost per sample and cost per data point (i.e., fractions of a cent per marker). The low-cost and high-throughput nature of GBS allows plant breeders to genotype thousands of plants per cycle in genomic selection driven breeding programs (He *et al.* 2014). Primary industry programmes in animal breeding have also taken up GBS. Unlike microsatellites or SNP chips, no previously generated genomic resources are necessary to deploy GBS. This allows researchers working in non-model species, such as orphan crops (i.e., crops of regional commercial importance, but not global), to take advantage of powerful genomic tools (Varshney *et al.* 2012). The situation for researchers in ecology and conservation biology is not dissimilar to that of those working with orphan crops. The budgets are small, resources meager and the questions are of local importance with small (if any) obvious economic returns. It is no wonder that ecologists were amongst the earliest adopters of GBS.

During the development of the GBS, we tested it on species other than maize. Confident that it worked in a variety of kingdoms, we welcomed interested early adopters to the lab for assistance. Two of those early adopters worked in the ecology space. Dr. Thomas White worked with the invasive bank vole (White *et al.* 2013) in Ireland which had small sample sizes and no reference genome. Dr. Nancy Chen studied the Florida scrub jay and developed a method using GBS data and Mendelian inheritance to improve SNP discovery (Chen *et al.* 2014). It became clear that we had developed a generally useful genomics research tool and it could be used by researchers across disciplines. We had already published the method in an open access journal and provided analysis software under a free software license. To allow researchers to more easily use this technology, we set up a GBS service at Cornell. By early 2016 the Cornell service had performed GBS analysis on over 1,500 species.

After our initial GBS publication, a plethora of method modifications and additional software tools have emerged. The recently published epiGBS method (VanGurp 2016) allows the interrogation of the methylome and does not require a reference genome, thereby extending the utility of the base method greatly. The GBSX toolkit (Herten *et al.* 2015) is a set of software designed to assist in the design of GBS based experiments. Many software packages have been developed to analyse GBS data (e.g., TASSEL-UNEAK, Stacks, GBS-SNP-CROPS, GbPSs; Lu *et al.* 2013; Catchen *et al.* 2013; Hapke & Thiele 2016; Melo *et al.* 2016) that are appropriate for species without reference genomes. Extensions to the molecular method and new software tools make these types of genomics approaches more broadly accessible; however, barriers to using this technology still exist in many disciplines, including the cost of laboratory and informatics setup and reservations in transitioning to new analytical tools.

Marker technology adoption has a long tail distribution. In 2013, I gave a talk on GBS at the *Molecular Markers in Horticulture Symposium*. Perusing the poster session, I found that researchers were using every type of marker technology that I knew about: from isozymes to GBS. Why were some researchers using cutting edge technologies? Why were others using antiquated, expensive and low information content technologies? Researchers in conservation genomics are in a similar situation. Across disciplines, the biological sciences are encountering rapidly changing technologies and increasingly larger data sets. Industry service providers with expert knowledge and experience, like my small New Zealand-based company (Elshire Group, Ltd.) and many others, can help bridge the gap. By developing relationships spanning human health, primary industry and conservation, as well as actively participating in research communities like MapNet (Box 3), we can work together to expedite the adoption of genomic technologies applicable to the questions at hand, effectively, efficiently and with confidence.

Relationships between conservation geneticists and primary industry scientists can result in improved commercial outcome for primary species as well. Conservation geneticists strive to preserve genetic diversity and the ecological and evolutionary processes that generate it (Groom *et al.* 2006; Haig *et al.* 2016). There is growing discussion among primary industry scientists regarding the need for commercial breeding programmes to maximise genetic diversity and minimise inbreeding (Medugorac *et al.* 2009; Windig & Engelsma 2010; Joost *et al.* 2011; Lenstra *et al.* 2012; Pryce *et al.* 2012; Kristensen *et al.* 2015). Livestock and crops are often of a small effective population size (i.e.,  $N_e < 100$ ) due to many generations of artificial selection for desired traits and are thus susceptible to loss of genome-wide variation via inbreeding and genetic drift (Windig & Engelsma 2010; Leroy *et al.* 2013; Kristensen *et al.* 2015; Jiménez-Mena *et al.* 2016; Shepherd *et al.* 2016). There is evidence for inbreeding depression in domestic species, such as cashmere goats (Dai *et al.* 2015), Iranian Guilan sheep (Eteqadi *et al.* 2015) and Iberian pigs (Saura *et al.* 2015). There is also an increasing awareness of the risks associated with deploying very few genotypes, particularly in the presence of novel crop pathogens (Kim *et al.* 2015) and an increasing concern among rare breeds regarding the loss of genetic variation associated with traits that might be useful in future markets (e.g., Catalanian donkey, Gutierréz *et al.* 2005; Famennoise poultry, Moula *et al.* 2009; black Slavonian pigs, Luković *et al.* 2012). Conservation geneticists have many of years of expertise regarding the conservation genetic management strategies for threatened species (Frankham 2010a). As a consequence, conservation geneticists can provide this biodiversity expertise to commercial species for improved primary production (Figure 3.2).

Conservation biologists and primary industry scientists also share similar goals regarding how best to mitigate the impact of climate change (Kristensen *et al.* 2015). For example, plant and animal breeders are prioritizing the selection of heat-tolerant plants (Ye *et al.* 2015) and low-emission animals (Hayes *et al.* 2013) and conservation scientists are debating a role for intentional introgression of desired phenotypic traits (e.g., heat tolerance) among locally adapted species or populations (Hamilton & Miller 2015; Kovach *et al.* 2016; Miller & Hamilton 2016). Given these shared goals, there is merit for scientists in primary industry and conservation to work together to maintain the evolutionary potential of commercial and threatened species in a changing climate.

A compelling rationale for building strong relationships between primary industry and conservation biology is that scientists in both disciplines conduct applied genetic research. Whereas primary industry scientists respond to the needs of primary industry practitioners (i.e., plant and animal breeders, farmers, fishermen and loggers), conservation scientists respond to the needs of conservation practitioners (i.e., wildlife managers and policy makers; Gordon *et al.* 2014; Haig *et al.* 2016). Considering the research-implementation gap that has been discussed in conservation genetic and genomic literature (Knight *et al.* 2008; Laikre *et al.* 2010; Shafer *et al.* 2015; Taylor & Soanes 2016), researchers from conservation genetics and primary industry can collaborate on how to best communicate research needs and results between scientists and practitioners. In the policy arena, both conservation geneticists and primary industry scientists work to develop improved policy regarding the utilisation and dissemination of genetic and genomic information (e.g., the *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization*, <https://www.cbd.int/abs/>; the *International Treaty on Plant Genetic Resources for Food and Agriculture*,

<http://www.planttreaty.org/>) and we anticipate that relationships between the two disciplines will allow for discussion on how to best form policy regarding the application of genomic information to threatened and commercial species.

Cross-sector collaborations will provide exciting opportunities to strategize how best to engage with stakeholders (e.g., private landowners, local governments, and research-funding bodies; Jacobson & Duff 1998; Dubbeling & Merzthal 2006); but where we see an even greater opportunity for considerable gains is for conservation geneticists and primary industry scientists to learn from one another about the importance of building meaningful partnerships with local and indigenous communities. Partnering with these communities enriches conservation and primary industry science because it creates research projects that are informed by the traditional knowledge and needs of these communities from the initial research proposal to the final report. In New Zealand, scientists and practitioners have clear directives to engage with Māori regarding the management of taonga (treasured) species (i.e., *Ko Aotearoa Tēnei/This is New Zealand*, conventionally known as WAI 262, <http://www.waitangitribunal.govt.nz/>) and various approaches have been developed to facilitate such engagement (Tipene-Matua & Henaghan 2007; Wilcox *et al.* 2008; Hudson *et al.* 2010). In addition, researchers are required to consult with relevant Māori tribes (iwi or hapu) when applying to receive permits for scientific research on taonga species from the Department of Conservation. New Zealand endemic species of cultural importance include threatened species (e.g., tuturuatu/shore plover and kakī/black silt; Box 1) and commercial species (e.g., pōrohe/green-lipped mussel, *Perna canaliculus*) and therefore we urge conservation genetic and primary industry scientists to collaborate on how to build productive partnerships with relevant Māori communities to develop research that is responsive to the needs and expectations of those communities. Beyond New Zealand,

researchers based in any of the 92 countries around the world that are signatories to the *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization* (<https://www.cbd.int/abs/>) have an opportunity to do the same. However, we argue that as global citizens, all scientists should be acting as if their country was a signatory, because as we get closer to generating population genomic datasets that include whole genomes for species of cultural importance we need to be more aware of how these genomic resources can affect and benefit local and indigenous communities.

### ***3.2.5: Moving Forward***

While multi-tasking empirical research, relationships with practitioners, stakeholders and interdisciplinary partnerships can be cumbersome, we are confident that the biggest gains in both conservation genetics and primary industry will be made under this approach. Given the mutual problems that can be solved when conservation geneticists and primary industry scientists work together, we encourage scientists in both disciplines to be leaders in interdisciplinary research and we offer the following advice on how to best forge these relationships:

#### **3.2.5.1: Get Out of Your Silo**

The first step to building successful interdisciplinary relationships is for researchers to get out of their silos and meet people with aligned research goals across disciplines. To accomplish this task for conservation genetics and primary industry, we advocate for small (<100 people) and diverse cross-sector meetings that allow participants from academia, government agencies and private institutions to actively engage with every presentation, especially those outside of their silos. In a New Zealand context, annual meetings such as

MapNet (see Box 3), the Canterbury 'Omics Symposium, and the Queenstown Research Week exemplify small, diverse, cross-sector meetings that allow scientists from both conservation and primary industry to meet and expand their research networks. For larger countries, these diverse and small meetings might be more effective on a regional versus a national level. In addition to meetings, we encourage conservation geneticists and primary industry scientists to attend genomic and networking workshops to meet people with aligned vision for genomic research, albeit in another discipline.

**Box 3. Building strong interdisciplinary relationships: MapNet and VISG. *Perspectives are those of Phil Wilcox.***

MapNet is a genomics collaboration that was formed in 2005 by a collective of New Zealand-based researchers from agriculture, horticulture, forestry and human medical genetics that quickly identified analytical gaps in international statistical genetics research. In response, MapNet members formed the Virtual Institute of Statistical Genetics (i.e., VISG) in 2007 and successfully obtained research funding to address these gaps. Through these synergies, methods developed for large human data sets (e.g., *CNVrd*, *CNVrd2*, *selectionTools*; Nguyen *et al.* 2013, 2014; Cadzow *et al.* 2014) have been successfully applied to apple data to identify genes of interest in commercial species (e.g., Boocock *et al.* 2015). Other workflows, such as the *selectionTools* pipeline developed and applied to human datasets such as the *1000Genomes* human data (Cadzow *et al.* 2014) are applicable to other outcrossed species where genetic maps are available. Recently, these relationships have also expanded to include cross-sector projects with scientists from the EEB and conservation genetics sector, who are able to provide insight into how these pipelines can be more broadly applicable to other applied genetic disciplines.

Critical for these cross-sectoral collaborations is effective and ethical behaviours among researchers, distributed leadership, commitment to an explicitly articulated vision, and effective resourcing for method development and testing. Ongoing cost reductions in both high-throughput sequencing and genotyping will constantly challenge data analyses. Thus collaborations among researchers in primary industry, human medical genetics, EEB and conservation genetics are an effective option to develop and apply genomic methods in a financially limited environment.

The benefits of the above-mentioned collaborations would ensure (a) relevant data analysis tools could be produced by adding relevance and utility to primary sector researchers proposing to develop such tools, and (b) providing a platform for more efficient utilisation of resources such as laboratory spaces and analytical capabilities, further reducing costs and therefore increasing data generation capacity. Collaborating with primary sector researchers working on closely related species would also benefit conservation genetics by improving efficiency. In some cases, the same species may be endangered within its natural range, but be of commercial value in other regions – such as *Pinus radiata*, which is widely planted as an exotic in the southern hemisphere but endangered in its natural range in Baja and northern California. An additional benefit of such collaborations is valuable experience and learnings from primary sector colleagues regarding experiment design, data analyses and interpretation of results. The MapNet collective was formed and run at essentially no cost, by utilising the resources of collaborating institutions and labour of those who were committed to this initiative, thus such cross-sector networks are easy to establish and operate – and often professionally rewarding for all involved.

### **3.2.5.2: Practice Leadership in Interdisciplinary Research**

The second step is to forge mutually beneficial partnerships between conservation and primary industry is to actively communicate with and collaborate with researchers

outside of one's silo. Doing so invariably requires leadership, respect and motivation to tackle shared problems (see Table 3.1), generally by expanding your own research programme to incorporate collaborative interdisciplinary projects between conservation and primary industry (e.g., Banks 2004; Knowler & Bradshaw 2007; Hobbs *et al.* 2008; Blank 2013; Sardinas & Kremen 2015; Box 3). Upon launching these collaborations, it is essential that leaders from both parties open an honest dialog concerning expectations, limitations, and potential hindrances to interdisciplinary work such as intellectual property issues. If collaborative groups choose to develop new methods or bioinformatic pipelines, we encourage these groups to test these tools on different species representing a wide-range of genomic complexities (i.e., ploidy levels, genome size, and number of repetitive elements, see Table 3.1) so these tools are robust and widely applicable to any research study (see also Box 2; Box 3). We also advocate for these collaborative groups to develop methods and pipelines that are open-source (see Box 2), which inspires others to use and improve upon cross-disciplinary tools. Pursuing co-funding opportunities between conservation and primary industry can be an excellent means of building mutually beneficial research collaborations, especially given that some grant providers favor collaborative proposals that tackle complex problems with broad research impact (Ledford 2015; but see also Bromham *et al.* 2016). Worldwide, there are groups that are forming to tackle complex problems through an interdisciplinary approach, including the Virtual Institute of Statistical Genetics (see Box 3) and Te Pūnaha Matatini (translated to “the meeting place of many faces”, <http://www.tepunahamatatini.ac.nz/>). As leaders from conservation and primary industry initialise interdisciplinary research, we encourage the formation and utilisation of these groups to facilitate the scientific process and encourage the involvement of new partners.



### 3.2.5.3: Promote a Community of Interdisciplinary Research

Leaders in both the conservation and primary industry sphere can go beyond collaborating with interdisciplinary scientists to promote a culture of interdisciplinary research. To accomplish this, we encourage editorial teams at conservation and ecology and evolution journals with a broad readership like *Molecular Ecology* to periodically invite perspective articles from colleagues in primary industry. We equate this approach to the recent decision made by the editorial team at *Animal Conservation* to invite submissions from conservation practitioners so conservation academics can better understand the needs and challenges of real-world conservation (Gordon *et al.* 2014). Leaders who are organising meetings and conferences in primary industry, conservation and genomics can strive to incorporate cross-sector talks and break down organisational silos by minimising field-specific sessions, as proposed by Taylor & Soanes (2016) and practiced by cross-sector meetings like MapNet (see Box 3). We also challenge scientists in both primary industry and conservation to become good interdisciplinary mentors to promote a culture of interdisciplinary research. This can involve mentors in conservation and primary industry promoting genomic seasonal internships or research positions to students in different silos. Not only will this encourage an interdisciplinary field, but it will also produce well-rounded and informed students with a network of colleagues to help solve shared problems.

After relationships between conservation genetics and primary industry are forged, we do not anticipate relationships will end once genomic gains are made in both disciplines. Instead, we envision these relationships will continue to grow and enable both disciplines to problem solve and incorporate new technologies for the improvement of threatened and commercial species. With other emerging techniques being discussed and used in both

conservation and primary industry, including other *-omic* techniques (e.g., transcriptomics, proteomics, metabolomics; Diz & Calvete 2016; Todd *et al.* 2016), epigenetic studies (Verhoeven *et al.* 2016) and genome editing (Johnson *et al.* 2016), we expect conservation genetics and primary industry to continue to collaborate and solve mutual problems while incorporating new technologies in an applied discipline.

We are confident that building strong interdisciplinary relationships will enable genomic advances in both conservation genetics and primary industry. However, we appreciate our colleagues in the global conservation community may be pursuing different strategies to successfully navigate the transition from genetics to genomics and we look forward to hearing about them in due course. In the meantime, our hope is that new technologies including genomics will be effectively incorporated into applied genetic disciplines like conservation and primary industry, because there is much to gain by using HTS technologies to improve outcomes for the world's threatened and commercial species.

### *3.2.6: Acknowledgements*

We appreciate the constructive insights from all anonymous reviewers of this article, particularly in regards to comments that helped us shape the action plan for interdisciplinary research between conservation genetics and primary industry. We thank Margaret Paterson, Mauricio Gonzales Chang and Ilina Cubrinovska for their assistance with the literature search. We are grateful to Paul Gardner for inadvertently inspiring this manuscript. We acknowledge Dr. Marty Faville for his contributions to our genetic and genomic cost comparisons. Finally, we thank Leonie Heyder for the tuturuatu image.

### *3.2.7: References*

- Ali OA, O'Rourke SM, Amish SJ, Meek MH, Luikart G, Jeffres C, Miller MR (2016) RAD Capture (Rapture): Flexible and Efficient Sequence-based genotyping. *Genetics*, **202**, 389-400.
- Allen AM, Barker GL, Wilkinson P, *et al.* (2012) Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, **11**, 279-295.
- Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.
- Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81-92.
- Angeloni F, Wagermaker N, Vergeer P, Ouborg J (2011) Genomic toolboxes for conservation biologists. *Evolutionary Applications*, **5**, 130–143.
- Aslam ML, Bastiaansen JWM, Elferink MG, *et al.* (2012) Whole genome SNP discovery and analysis of genetic diversity in Turkey (*Melagris gallopavo*). *BMC Genomics*, **13**, 391.
- Aulchenko S, de Koning D, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577–585.
- Avise JC (2008) The history, purview, and future of conservation genetics. In: Conservation biology: evolution in action. Oxford University Press: 5-15.
- Avise, JC (2010) Perspective: conservation genetics enters the genomics era. *Conservation Genetics*, **11**, 665-669.
- Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.
- Ball RD (2005) Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Genetics*, **170**, 859-873.
- Banks JE (2004) Divided culture: integrating agriculture and conservation biology. *Frontiers in Ecology and the Environment*, **2**, 537-545.
- Bataillon TM, David JL, Schoen DJ (1996) Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, **144**, 409-417.
- Begum H, Spindel JE, Lalusin A, *et al.* (2015) Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*) *PLoS ONE*, **10**, e0119873.

- Benestan LM, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, **24**, 3299-3315.
- Benestan LM, Ferchaud L, Hohenlohe PA, Garner BA, *et al.* (2016) Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular Ecology*, **25**, 2967–2977.
- Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2016) Genomic analysis reveals depression due to both individual and maternal inbreeding in a free-living mammal population. *Molecular Ecology*, **25**, 3152-3168.
- Bertioli DJ, Ozias-Akins P, Chu Y, *et al.* (2014) The Use of SNP Markers for Linkage Mapping in Diploid and Tetraploid Peanuts. *G3*, **4**, 89-96.
- Blank PJ (2013) Northern bobwhite response to conservation reserve program habitat and landscape attributes. *The Journal of Wildlife Management*, **77**, 68-74.
- Boocock J, Chagne D, Merriman TR, Black MA (2015) The distribution and impact of common copy-number variation in the genome of the domesticated apple, *Malus x domestica* Borkh. *BMC Genomics*, **16**, 848.
- Bowden R, MacFie TS, Myers S, *et al.* (2012) Genomic tools for evolution and conservation in the chimpanzee: *Pan troglodytes ellioti* is a genetically distinct population. *PLoS Genetics*, **8**, e1002504.
- Brachi B, Meyer CG, Villoutreix R, Platt A, Mortona TC, Roux F, Bergelson J (2015) Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proceedings of the National Academy of Science*, **112**, 4032-4037.
- Bromham L, Dinnage R, Hua X (2016) Interdisciplinary research has consistently lower funding success. *Nature*, **534**, 684-687.
- Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA (2012) Development and mapping of SNP assays in allotetraploid cotton. *Theoretical Applied Genetics*, **124**, 1201-1214.
- Caballero A, García-Dorado (2013) Allelic diversity and its implications for the rate of adaptation. *Genetics*, **195**, 1373-1384.
- Cadzow M, Boocock J, Nguyen HT, Wilcox P, Merriman TR, Black MA (2014) A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics*, **5**, 293.
- Card DC, Schield DR, Reyes-Velasco J, *et al.* (2014) Two low coverage bird genomes and a comparison of reference-guided versus *de novo* genome assemblies. *PLoS ONE*, **9**, e106649.

- Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124-3140.
- Chaves JA, Cooper EA, Hendry AP, *et al.* (2016) Genomic variation at the tips of the adaptive radiation of Darwin's finches. *Molecular Ecology*, **25(21)**, 5282-5295.
- Chen N, van Hout CV, Gottipati S, Clari AG (2014) Using Mendelian Inheritance To Improve High-Throughput SNP Discovery. *Genetics*, **198**, 847-857.
- Clevenger J, Chararro C, Pearl SA, Ozias-Akins, Jackson SA (2015) Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. *Molecular Plant*, **8**, 831-846.
- Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B*, **363**, 557–572.
- Cubrinovska I, Massaro M, Hale ML (2016) Assessment of hybridisation between the endangered Chatham Island black robin (*Petroica traversi*) and the Chatham Island tomtit (*Petroica macrocephala chathamensis*). *Conservation Genetics*, **17**, 259-265.
- da Silva AG, Appleyard SA, Upston J (2015) Establishing the evolutionary compatibility of potential sources of colonizers for overfished stocks: a population genomics approach. *Molecular Ecology*, **24**, 564-579.
- Dai S, Wang C, Wang Z, *et al.* (2015) Inbreeding and its effects on fleece traits of inner Mongolia cashmere goats. *Small Ruminant Research*, **128**, 50-53.
- Defaveri J, Viitaniemi H, Leder E, Merilä (2013) Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. *Molecular Ecology Resources*, **13**, 377-392.
- Dierickx EG, Shultz AJ, Sato F, Hiraoka T, Edwards SV (2015) Morphological and genomic comparisons of Hawaiian and Japanese Black-footed Albatrosses (*Phoebastria nigripes*) using double digest RADseq: implications for conservation. *Evolutionary Applications*, **8**, 662-678.
- Dillon SK, Nolan MF, Matter P, Gapare WJ, Bragg JG, Southerton SG (2013) Signatures of adaptation and genetic structure among the mainland populations of *Pinus radiata* (D. Don) inferred from SNP loci. *Tree Genetics & Genomes*, **9**, 1447-1463.
- Dodds KG, McEwan JC, Brauning R, Anderson RM, van Stijn TC, Kristjánsson T, Clarke SM (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, **16(1)**, 1047.

- Diz AP, Calvete JJ (2016) Ecological proteomics: is the field ripe for integrating proteomics into evolutionary ecology research? *Journal of Proteomics*, **135**, 1-3.
- Dowling TE, Anderson CD, Marsh PC, Rosenberg MS (2015) Population Structure in the Roundtail Chub (*Gila robusta* Complex) of the Gila River Basin as Determined by Microsatellites: Evolutionary and Conservation Implications. *PLoS ONE*, **10**, e0139832.
- Dubbeling M, Merzthal G (2006) Sustaining urban agriculture requires the involvement of multiple stakeholders. In: *Cities farming for future, Urban Agriculture for green and productive cities*. IDRC and IIRR Publishing: 20-51.
- Edmands S (2007) Between a rock and a hard place: evaluating the relative risks of inbreeding and outbreeding depression for conservation and management. *Molecular Ecology*, **16**, 463-475.
- Eklom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1-15.
- Ellegren, H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, **29**, 51-63.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, **6**, e19379.
- Eteqadi B, Hossein-Zadeh NG, Shadparvar AA (2015) Inbreeding effects on reproductive traits in Iranian Guilan sheep. *Tropical Animal Health and Production*, **47**, 533-539.
- Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*, 4th ed. Edition. Longman, Essex.
- Frankham R (1995) Conservation genetics. *Annual Review of Genetics*, **29**, 305-327.
- Frankham R (2005) Genetics and extinction. *Biological Conservation*, **126**, 131-140.
- Frankham R. 2010a. Where are we in conservation genetics and where do we need to go? *Conservation Genetics*, **11**, 661-663.
- Frankham R. 2010b. Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, **143**, 1919-1927.
- Frankham R, Ballou JD, Eldridge MD, Lacy RD, Ralls K, Dudash MR, Fenster CB (2011) Predicting the probability of outbreeding depression. *Conservation Biology*, **25**, 265-275.

- Frankham R (2015) Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Molecular Ecology*, **24**, 2610-2608.
- Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology and Evolution*, **27**, 489-496.
- Garner BA, Hand BK, Amish SJ, *et al.* (2016) Genomics in Conservation: Case Studies and Bridging the Gap Between Data and Application. *Trends in Ecology and Evolution*, **31**, 81-83.
- George AW, Visscher PM, CS Haley (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, **156**, 2081-2092.
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, e90346.
- Goff SA, Ricke D, Lan TH, *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. Japonica). *Science*, **296**, 92-100.
- Gordon IJ, Evans DM, Garnder TWJ, *et al.* (2014) Enhancing communication between conservation biologists and conservation practitioners: letter from the conservation front line. *Animal Conservation*, **17**, 1-2.
- Gore MA, Chia JM, Elshire RJ, *et al.* (2009) A First-Generation Haplotype Map of Maize. *Science*, **326**, 1115-1117.
- Groom MJ, Meffe GK, Carroll CR (2006). *Principles of Conservation Biology* (No. 333.9516 G7). Sinauer Associates, Sunderland.
- Grueber CE (2016) Comparative genomics for biodiversity conservation. *Computational and Structural Biotechnology*, **13**, 370-375.
- Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nature*, **37**, 549-554.
- Gutiérrez JP, Marmi J, Goyache F, Jordana J (2005) Pedigree information reveals moderate to high levels of inbreeding and a weak population structure in the endangered Catalanian donkey breed. *Journal of Animal Breed Genetics*, **122**, 378-386.
- Hagen EN, Hale ML, Maloney RF, Steeves TE (2011) Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the Black Stilt *Himantopus novaezelandiae*. *Ibis*, **153**, 556-561.
- Haig SM, Miller MP, Bellinger R, Draheim HM, Mercer DM, Mullins TD (2016) The conservation genetics juggling act: integrating genetics and ecology, science and policy. *Evolutionary Applications*, **9**, 181–195.

- Halbert ND, Ward TJ, Schnabel RD, Taylor JF, Derr JN (2005) Conservation genomics: disequilibrium mapping of domestic cattle chromosomal segments in North American bison populations. *Molecular Ecology*, **14**, 2343-2362.
- Hammerly SC, de la Cerda DA, Bailey J, Johnson JA (2016) A pedigree gone bad: increased offspring survival after using DNA-based relatedness to minimize inbreeding in a captive population. *Animal Conservation*, **19**, 296-303.
- Hamilton JA, Miller JM (2015) Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology*, **30**, 33-41.
- Hapke A, Thiele D (2016) GIBPSs: a toolkit for fast and accurate analyses of genotyping-by-sequencing data without a reference genome. *Molecular Ecology Resources*, **16**, 979-990.
- Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014) Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, **7**, 1008–1025.
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, **92**, 433–443.
- Hayes BJ (2011) Technical note: efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. *Journal of Dairy Science*, **94**, 2114-2117.
- Hayes BJ, Lewin JA, Goddard ME (2013) The future of livestock breeding: Genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends in Genetics*, **29**, 206-214.
- He J, Zhao X, Laroche A, Lu Z, Liu H, Li Z (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, **5**, 484.
- He X, Johansson ML, Heath DD (2016) Role of genomics and transcriptomics in selection of reintroduction source populations. *Conservation Biology*, **30**(5), 1010-1018.
- Heffner EL, Sorrells ME, Jannink J (2009) Genomic selection for crop improvement. *Crop Science*, **49**, 1-12.
- Herrero-Medrano JM, Megens HJ, Crooijmans RP, Abellaneda JM, Ramis G (2012) Farm-by-farm analysis of microsatellite, mtDNA, and SNP genotype data reveals inbreeding and crossbreeding as threats to the survival of a native Spanish pig breed. *Animal Genetics*, **44**, 259-66.



- Herten K, Hestand MS, Vermeesch JR, Van Houdt JKJ (2015) GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics*, **16**, 73.
- Hess JE, Zendt JS, Matala AR, Narum SR (2016) Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. *Proceedings of the Royal Society B*, **283**, 20153064.
- Hillier L, Miller W, Birney E, *et al* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695-716.
- Hobbs PR., Sayre K, Gupta R (2008) The role of conservation agriculture in sustainable agriculture. *Philosophical Transactions of the Royal Society Biology*, **363**, 543-555.
- Hoffman JL, Simpson F, David P, *et al.* (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Science*, **111**, 3775-3780.
- Hogg C, Grueber CE, Pemberton D, Fox S, Lee AV, Ivy JA, Velov K (2016) “Devil Tools & Tech”: a synergy of conservation research and management practices. *Conservation Letters*, **10**(1), 133-138.
- Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117-122.
- Hu ZL, Park CA, Wu XL, Reecy HM (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Research*, **41**, 871–879.
- Hudson M, Milne M, Reynolds P, Russell K, Smith B (2010) Te Ara Tika Guidelines for Māori research ethics: A Framework for Researchers and Ethics Committee Members. ISBN: 978-1-877495-03-8.
- IUCN/SSC (2013) Guidelines for Reintroductions and Other Conservation Translocations. Version 1.0. IUCN Species Survival Commission, Gland Switzerland; Cambridge, UK.
- Jacobson SK, Duff MD (1998) Training idiot savants: the lack of human dimensions in conservation biology. *Conservation Biology*, **12**, 263-267.
- Jiménez-Mena B, Hospital F, Bataillon T (2016) Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. *Conservation Genetics Resources*, **8**, 35-41.
- Johnson JA, Altwegg R, Evans DM, Ewen JG, Gordon IJ, Pettorelli N, Young JK (2016) Is there a future for genome-editing technologies in conservation? *Animal Conservation*, **19**, 97-101.

- Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185-202.
- Joost S, Colli L, Bonin A, Beibach I, Allendorf FW, *et al.* (2011) Promoting collaboration between livestock and wildlife conservation genetic communities. *Conservation Genetic Resources*, **3**, 785-788.
- Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, **115**, 63-72.
- Keller LF, Arcese P, Smith JNM, Hochachka WM, Stearns SC (1994) Selection against inbred song sparrows during a natural population bottleneck. *Nature*, **372**, 356-357.
- Kim HB, Joa JH, Song KJ (2015) Current status and prospects of kiwifruit (*Actinidia chinensis*) genomics. *Journal of Plant Biotechnology*, **42**, 342-349.
- Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, Forstmeier W (2015) Quantifying realized inbreeding in wild and captive animal populations. *Heredity*, **114**, 397-403.
- Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. *Trends in Ecology and Evolution*, **21**, 629-637.
- Knight AT, Cowling RM, Rouget M, Balmford A, Lombard AT, Campbell BM (2008) Knowing but not doing: selecting priority conservation areas and the research-implementation gap. *Conservation Biology*, **22**, 610-617.
- Knowler D, Bradshaw B (2007) Farmer's adoption of conservation agriculture: A review and synthesis of recent research. *Food Policy*, **32**, 25-48.
- Kovach RP, Luikart G, Lower WH, Boyer MC, Muhlfeld CC (2016) Risk and efficacy of human-enabled interspecific hybridization for climate-change adaptation: response to Hamilton and Miller (2016). *Conservation Biology*, **30**, 428-430.
- Kraus RH, Vonholdt B, Cocchiara B, *et al.* (2015). A single-nucleotide polymorphism-based approach for rapid and cost-effective genetic wolf monitoring in Europe based on noninvasively collected samples. *Molecular ecology resources*, **15**, 295-305.
- Kristensen TN, Hoffmann AA, Pertoldi C, Stronen AV (2015) What can livestock breeders learn from conservation genetics and vice versa? *Frontiers in Genetics*, **6**, 1-12.
- Laikre L, Allendorf FW, Aroner LC, *et al.* (2010) Neglect of genetic diversity in implementation of the convention on biological diversity. *Conservation Biology*, **24**, 86-88.
- Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE (2014) Genotyping by sequencing resolves shallow population structure to inform

- conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, **7**, 355-369.
- Ledford H (2015) Team Science. *Nature*, **525**, 308-311.
- Lemay MA, Russello MA (2015) Genetic evidence for ecological divergence in kokanee salmon. *Molecular ecology*, **24**, 798-811.
- Lenstra JA, Groeneveld LF, Eding J, *et al.* (2012) Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. *Animal Genetics*, **43**, 483–502.
- Leroy G, Mary-Huard T, Danvy S, Charvolin E, Danchin-Burge C (2013) Methods to estimate effective population size using pedigree data: examples in dog, sheep, cattle, and horse. *Genetics Selection Evolution*, **45**, 1.
- Lew RM, Finger AJ, Baerwald MR, Goodbla A, May B, Meek MH (2015) Using next-generation sequencing to assist a conservation hatchery: a single-nucleotide polymorphism panel for the genetic management of endangered delta smelt. *Transactions of the American Fisheries Society*, **144**, 767-779.
- Li M, Strandén I, Tiirikka T, Sevón-Aimonen M, Kantanen J (2011) A comparison of approaches to estimate the inbreeding coefficient and pairwise relatedness using genomic and pedigree data in a sheep population. *PLoS ONE*, **6**, e26256.
- Li S, Li B, Cheng C, *et al.* (2014) Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biology*, **15**, 557.
- Li Y, Lancaster ML, Cooper SJB, Taylor AC, Carthew SM (2015a) Population structure and gene flow in the endangered southern brown bandicoot (*Isodon obesulus obesulus*) across a fragmented landscape. *Conservation Genetics*, **16**, 331-345.
- Li X, Buitenhuis AJ, Lund MS, *et al.* (2015b) Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. *Journal of Dairy Science*, **98**, 8152-8163.
- Ljungqvist M, Åkesson M, Hansson B (2010) Do microsatellites reflect genome-wide genetic diversity in natural populations? A comment on Väli *et al.* (2008) *Molecular Ecology*, **19**, 851–855.
- Lu F, Lipka AE, Glaubitz J, *et al.* (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genetics*, **9**, e1003215.
- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

- Luković Z, Karolyi D, Klišanić V, Mahnet Ž, Škorput D (2012) Genetic parameters and trends for litter size in Black Slavonian pigs. *7th International Symposium on the Mediterranean Pig*, ISBN: 2-85352-488-4.
- Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland.
- Margulies M, Egholm M, Altman WE, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.
- Marshall C, Chagné D, Deusch O, *et al.* (2015) A DNA-based diagnostic for differentiating among New Zealand endemic *Podocarpus*. *Tree Genetics & Genomes*, **11**, 1-13.
- Matukumalli LK, Lawley CT, Schnabel RD, *et al.* (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE*, **4**, e5350.
- McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526-538.
- McMahon BJ, Teeling EC, Höglund J (2014) How and why should we implement genomics into conservation? *Evolutionary Applications*, **7**, 999–1007.
- Medugorac I, Medugorac A, Russ I, *et al.* (2009) Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. *Molecular Ecology*, **18**, 3394-3410.
- Melo ATO, Bartaula R, Hale I (2016) GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics*, **17**, 29.
- Miller JM, Poissant J, Hogg JT, Coltman DW (2012) Genomic consequences of genetic rescue in an insular population of bighorn sheep (*Ovis canadensis*). *Molecular Ecology*, **21**, 1583-1596.
- Miller JM, Hamilton JA (2016) Interspecies hybridization in the conservation toolbox: response to Kovach *et al.* (2016). *Conservation Biology*, **30**, 431-433.
- Moula N, Antoine-Moussiaux N, Farnir F, Leroy P (2009) Evaluation of the production performances of an endangered local poultry breed, the famennoise. *International Journal of Poultry Science*, **8**, 389-396.
- Muir WM, Wong GKS, Zhang Y, *et al.* (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Science*, **105**, 17312-17317.

- Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841-2847.
- Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, **12**, 111-112.
- Nguyen HT, Merrimen TR, Black MA (2013) CNVrd, a Read-Depth Algorithm for Assigning Copy-Number at the FCGR Locus: Population-Specific Tagging of Copy Number Variation at *FCGR3B*. *PLoS ONE*, **8**, e63219.
- Nguyen HT, Merrimen TR, Black MA (2014) The CNVrd2 package: measurement of copy number at complex loci using high-throughput sequencing data. *Frontiers in Genetics*, **5**, 248.
- Ogden R, Baird J, Senn H, McEwing R (2012) The use of cross-species genome-wide arrays to discover SNP markers for conservation genetics: a case study from Arabian and scimitar-horned oryx. *Conservation Genetic Resources*, **4**, 471-473.
- Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics*, **26**, 177-187.
- Overbeek AL, Hauber ME, Brown E, Cleland S, Maloney RF, Steeves TE (2017) Evidence for brood parasitism in a critically endangered Charadriiform with implications for conservation. *Journal of Ornithology*, **158**, 333-337.
- Oyler-McCance SJ, Cornman RS, Jones KL, Fike JA (2015) Genomic single-nucleotide polymorphisms confirm that Gunnison and Greater sage-grouse are genetically well differentiated and that the bi-state population is distinct. *Condor*, **117**, 217-227.
- Pacioni C, Hunt H, Allentoft ME, Vaughan TG, Wayne AF, Baynes A, Haouchar D, Dortch J, Bunce M (2015) Genetic diversity loss in a biodiversity hotspot: ancient DNA quantifies genetic decline and former connectivity in a critically endangered marsupial. *Molecular Ecology*, **24**, 5813-5828.
- Pardo-Diaz C, Salazar C, Jiggins CD (2015) Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, **6**, 445-464.
- Payseur BA, Rieseberg LH (2016) A genomic perspective on hybridization and speciation. *Molecular Ecology*, **25**, 2337-2360.
- Pertoldi C, Wójcik JM, Tokarska M, *et al.* (2009) Genome variability in European and American bison detected using the BovineSNP50 bead chip. *Conservation Genetics*, **11**, 627-634.
- Poissant J, Shafer BA, Davis S, Mainguy J, Hogg JT, Côté SD, Coltman DW (2009) Genome-wide cross-amplification of domestic sheep microsatellites in bighorn sheep and mountain goats. *Molecular Ecology Resources*, **9**, 1121-1126.

- Price, A.L., N.A. Zaitlen, D. Reich & N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. *Nature Review Genetics*, **11**, 459–463.
- Primmer CR (2009) From conservation genetics to conservation genomics. *Annals of the New York Academy of Science*, **1162**, 357–368.
- Prince DJ, O'Rourke SM, Thompson TQ, *et al.* (2017) The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *Science Advances*, **3(8)**, e1603198.
- Pryce JE, Hayes BJ, Goodard ME (2006) Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *Journal of Dairy Science*, **95**, 377-388.
- Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RAD seq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, **e431**.
- Rietveld CA, Medland SE, Derringer J, *et al.* (2013) GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science*, **340**, 1467-1471.
- Robinson JA, Vecchy DO, Fan Z (2016) Genomic flatlining in the endangered island fox. *Current Biology*, **26**, 1183-1189.
- Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: Empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.
- Sardinas HX, Kremen C (2015) Pollination services from field-scale agricultural diversification may be context-dependent. *Agriculture, Ecosystems and Environment*, **207**, 17-25.
- Saura M, Fernández A, Varona L, *et al.* (2015) Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data. *Genetics Selection Evolution*, **47(1)**, 1.
- Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807-820.
- Schibler L, Roig A, Mahe MF, *et al.* (2004) A first generation bovine BAC-based physical map. *Genetics selection evolution*, **36**, 105-122.
- Schielzeth H, Husby A (2014) Challenges and prospects in genome-wide quantitative trait loci mapping of standing genetic variation in natural populations. *Annals of the New York Academy of Sciences*, **1320**, 35-57.
- Seddon PJ (2010) From Reintroduction to Assisted Colonization: Moving along the Conservation Translocation Spectrum. *Restoration Ecology*, **18**, 796-802.

- Sellner EM, Kim JW, McClure MC, Taylor KH, Schnabel RD, Taylor JF (2007) Board-invited review: applications of genomic information in livestock. *Journal of Animal Science*, **85**, 3148-3158.
- Shafer ABA., Wolf JBW, Alves PC, *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, **30**, 78-87.
- Shafer ABA, Wolf JBW, Alves PC, *et al.* (2016) Reply to Garner *et al.* *Trends in Ecology and Evolution*, **31**, 83.
- Shepherd LD, de Lange PJ, Cox S, McLenachan PA, Roskrige NR, Lockhart PJ (2016) Evidence of a strong domestication bottleneck in the recently cultivated New Zealand endemic root crop, *Arthropodium cirratum* (Asparagaceae). *PLOS One*, **11**(3), e0152455.
- Silió L, Barragán C, Fernández AI, García-Casco J, Rodríguez MC (2016) Assessing effective population size, coancestry and inbreeding effects on litter size using the pedigree and SNP data in closed lines of the Iberian pig breed. *Journal of Animal Breeding and Genetics*, **133**, 145-154.
- Slate J, Santure AW, Feulner PGD, Brown EA, Ball AD, Johnston SE, Gratten J (2010) Genome mapping in intensively studied wild vertebrate populations. *Trends in Ecology and Evolution*, **26**, 275-284.
- Sonah H, Deshmukh RK, Singh VP, Gupta DK, Singh NK, Sharma TR (2011) Genomic resources in horticultural crops: Status, utility and challenges. *Biotechnology Advances*, **29**, 199-209.
- Sovic MG, Fries AC, Gibbs HL (2015) AftRAD: a pipeline for accurate and efficient de novo assembly of RADseq data. *Molecular Ecology Resources*, **15**, 1163-1171.
- Soulé ME (1985) What is conservation biology? A new synthetic discipline addresses the dynamics and problems of perturbed species, communities, and ecosystems. *Bioscience*, **35**, 727-734.
- Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ecblof R, Bennison C, Ball AD, Beckerman AP, Slate J (2010) Adaptation genomics: the next generation. *Trends in Ecology and Evolution*, **25**, 705-712.
- Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL (2006) Whole-genome genotyping with the single-base extension assay. *Nature Methods*, **3**, 31-33.
- Steeves TE, Maloney RF, Hale ML, Tylianakis JM, Gemmell NJ (2010) Genetic analyses reveal hybridization but no hybrid swarm in one of the world's rarest birds. *Molecular Ecology*, **19**, 5090-5100.
- Steiner CC, Putnam AS, Hoeck PEA, Ryder OA (2013) Conservation genomics of threatened animal species. *Annual Review of Animal Bioscience*, **1**, 261-281.

- Taylor HR, Kardos MD, Ramstad KM, Allendorf FW (2015) Valid estimates of individual inbreeding coefficients from marker-based pedigrees are not feasible in wild populations with low allelic diversity. *Conservation Genetics*, **16**, 901-913.
- Taylor HR, Soanes K (2016) Breaking out of the echo chamber: missed opportunities for genetics at conservation conferences. *Biodiversity and Conservation*, **25(10)**, 1987-1993.
- Tipene-Matua B, Henaghan M (2007) Establishing a Māori ethical framework for genetic research with Māori. In: *Genes, Society and the Future* (ed Henaghan M), pp. 1-44. Human Genome Research Project, Dunedin, New Zealand.
- Tigano A, Friesen VL (2016) Genomics of local adaptation with gene flow. *Molecular Ecology*, **25**, 2144-2164.
- Todd EV, Black MA, Gemmell NJ (2016) The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology*, **25**, 1224-1241.
- Trask AE, Signal EM, McCracken DI, Monaghan P, Piertney SB, Reid JM (2015) Evidence of the phenotypic expression of a lethal recessive allele under inbreeding in a wild population of conservation concern. *Journal of Animal Ecology*, **85**, 879-891.
- Tuberosa R, Salvi S (2006) Genomics-based approaches to improve drought tolerance of crops. *Trends in Plant Science*, **11**, 405-412.
- Uitdewilligen JGAML, Wolters AA, D'hoop BB, Borm TJA, Visser RGF, VanEck HJ (2013) A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLoS*, **10**, e0141940.
- Väli Õ, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, **17**, 3808-3817.
- Varshney RK, Ribaut J, Buckler ES, Tuberosa R, Rafalski JA, Langridge P (2012) Can genomics boost productivity of orphan crops? *Nature Biotechnology*, **30**, 1172-1176.
- Verhoeven KJF, vonHoldt BM, Sork VL (2016) Epigenetics in ecology and evolution: what we know and what we need to know. *Molecular Ecology*, **25**, 1631-1638.
- Vilas A, Pérez-Figueroa A, Quesada H, Caballero A (2015) Allelic diversity for neutral markers retains a higher adaptive potential for quantitative traits than expected heterozygosity. *Molecular Ecology*, **24**, 4419-4432.
- Wang B, Ekblom R, Bunikis I, Siitari H, Höglund J (2014) Whole genome sequencing of the black grouse (*Tetrao tetrix*): reference guided assembly suggests faster-Z and MHC evolution. *BMC Genomics*, **15**, 180.



- Wang L, Zu L, Liu X (2015) Copy number variation-based genome wide association study reveals additional variants contributing to meat quality in swine. *Scientific Reports*, **5**, DOI:10.1038/srep12535.
- Weeks AJ, Sgro CM, Young AG, *et al.* (2011) Assessing the benefits and risks of translocations in changing environments: a genetic perspective. *Evolutionary Applications*, **4**, 709-725.
- White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. *Molecular Ecology*, **22**, 2971-2985.
- Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA (2015). Genetic rescue to the rescue. *Trends in Ecology & Evolution*, **30**, 42-49.
- Wilcox PL, Charity JA, Roberts MR, *et al.* (2008) A values-based process for cross-cultural dialogue between scientists and Māori. *Journal of the Royal Society of New Zealand*, **38**, 215-227.
- Willoughby JR, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, DeWoody JA (2015) The impacts of inbreeding, drift and selection on genetic diversity in conservation breeding populations. *Molecular Ecology*, **24**, 98-110.
- Windig JJ, Engelsma KA (2010) Perspectives of genomics for genetic conservation of livestock. *Conservation Genetics*, **11**, 635-641.
- Womack JE (2005) Advances in livestock genomics: opening the barn door. *Genome Research*, **15**, 1699-1705.
- Wuerschum T, Boeven PHG, Langer SM, Longin CFJ, Leiser W (2015) Multiply to conquer: copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genetics*, **16**, 96.
- Xia Q, Zhou S, Lu C, *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937-1940.
- Ye C, Tenorio F, Argayoso MA, Laza MA, Koh HJ, Redona ED, Jagadish KSV, Gregorio GB (2015) Identifying and confirming quantitative trait loci associated with heat tolerance at flowering stage in different rice populations. *BMC Genetics*, **16**, DOI: 10.1186/s12863-015-0199-7.
- Zhang G, Li C, Li Q, *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311-1320.
- Zhang C, Wang Z, Bruce H, Kem RA, Charagu P, Miar Y, Yang T, Plastow G (2015) Genome-wide association studies (GWAS) identify a QTL close to PRKAG3 affecting meat pH and colour in crossbred commercial pigs. *BMC Genetics*, **16**, DOI: 10.1186/s12863-015-0192-1.

Zhou Z, Jiang Y, Wang Z, *et al.* (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, **33**, 408-413.

### *3.2.8: Data accessibility*

Text files including all literature search results presented in Figure 3.1 and Supplemental Figure 3. 1 are available on Dryad: DOI:10.5061/dryad.32j55

# Chapter 4: Adding Tools to the Conservation Genomics Toolbox

## 4.1: Preamble

**Chapter 3** of this thesis addresses the conservation genomics gap and provides an approach for upskilling in bioinformatic expertise through building mutually-beneficial collaborations with primary industry research. Another contributing factor to the conservation genomics gap is a lack of readily available genomic resources for non-model threatened species. In **Chapter 4** — in a manuscript published in the journal *Genes* in 2019— my co-authors and I use genome resources developed in partnership with researchers at the University of Otago (Natalie Forsdick and Michael Knapp), the Institute of Clinical Molecular Biology at Kiel University (Marc Hoepfner), and Tea Break Bioinformatics Limited (Roger Moraga) to test whether SNPs discovered mapping kakī reads (i.e., genotyping-by-sequencing and whole genome resequencing) to reference genomes of closely related species will result in similar diversity estimates that could be useful for conservation management.

The published version of this chapter can be found in **Appendix D**, with the Supplemental Materials for this chapter found in **Appendix E**, and all scripts for SNP discovery, filtering, and analysis found in **Appendix F**.

### 4.1.1: Contribution Statement

This piece was co-first authored by myself and Natalie Forsdick (PhD Candidate, University of Otago), with project conceptualisation by all co-authors. Reference genomes shown here were generated and assembled by Natalie Forsdick (kakī and pied stilt), Marc

Hoepfner & Roger Moraga (pied avocet), and Roger Moraga (killdeer, with raw reads from the Birds 10,000 Genomes (B10K) Project; Zhang *et al.* 2014). I extracted DNA and performed quality control measures for reduced representation and resequencing samples shown here. Library preparation and sequencing was performed by The Elshire Group Limited (kakī genotyping-by-sequencing) and the Institute for Clinical Molecular Biology at Kiel University (kakī resequencing). I led SNP discovery using reduced representation and resequencing approaches across all reference genomes used, with guidance from Roger Moraga. All diversity estimates statistical analyses presented here were produced by myself. I led the first and final drafts of the manuscript, with assistance from my co-first author Natalie Forsdick and feedback from all coauthors. All figures seen here were developed and produced by myself, with tables co-produced by myself and Natalie Forsdick.



**Stephanie J. Galla**

## 4.2: Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management

Stephanie J. Galla <sup>1,\*</sup>, Natalie J. Forsdick <sup>2,\*</sup>, Liz Brown <sup>3</sup>, Marc P. Hoepfner <sup>4</sup>, Michael Knapp <sup>2</sup>, Richard F. Maloney <sup>5</sup>, Roger Moraga <sup>6</sup>, Anna W. Santure <sup>7</sup> and Tammy E. Steeves <sup>1</sup>

<sup>1</sup>School of Biological Sciences, University of Canterbury, Christchurch 8140, New Zealand

<sup>2</sup>Department of Anatomy, University of Otago, Dunedin 9054, New Zealand

<sup>3</sup>Te Manahuna, Department of Conservation, Twizel 7901, New Zealand

<sup>4</sup>Institute of Clinical Molecular Biology, Kiel University, Kiel 24105, Germany

<sup>5</sup>Science and Policy Group, Department of Conservation, Christchurch 8011, New Zealand

<sup>6</sup>Tea Break Bioinformatics Ltd., Palmerston North, 4144, New Zealand

<sup>7</sup>School of Biological Sciences, University of Auckland, Auckland 1142, New Zealand

\* Correspondence: stephanie.galla@pg.canterbury.ac.nz (S.J.G.);  
natalie.forsdick@postgrad.otago.ac.nz (N.J.F.)

#### 4.2.1: Abstract

Threatened species recovery programmes benefit from incorporating genomic data into conservation management strategies to enhance species recovery. However, a lack of readily available genomic resources, including conspecific reference genomes, often limits the inclusion of genomic data. Here, we investigate the utility of closely related high-quality reference genomes for single nucleotide polymorphism (SNP) discovery using the critically endangered kakī/black stilt (*Himantopus novaezelandiae*) and four Charadriiform reference genomes as proof of concept. We compare diversity estimates (i.e., nucleotide diversity, individual heterozygosity, and relatedness) based on kakī SNPs discovered from genotyping-by-sequencing and whole genome resequencing reads mapped to conordinal (killdeer, *Charadrius vociferus*), confamilial (pied avocet, *Recurvirostra avosetta*), congeneric (pied stilt, *Himantopus himantopus*) and conspecific reference genomes. Results indicate that diversity estimates calculated from SNPs discovered using closely related reference genomes correlate significantly with estimates calculated from SNPs discovered using a conspecific genome. Congeneric and confamilial references provide higher correlations and more similar measures of nucleotide diversity, individual heterozygosity, and relatedness. While conspecific genomes may be necessary to address other questions in conservation, SNP discovery using high-quality reference genomes of closely related species is a cost-effective approach for estimating diversity measures in threatened species.

**Keywords:** conservation genomics; conservation genomics gap; SNP discovery; B10K; threatened species; birds

#### 4.2.2: Introduction

The field of conservation genetics is in transition from using relatively few genetic markers (e.g., microsatellites, mitochondrial sequences) to using thousands of genome-wide single nucleotide polymorphisms (SNPs) discovered with high-throughput sequencing technologies (HTS) to inform conservation management of threatened species. In addition to providing greater resolution for diversity estimates (e.g., nucleotide diversity, heterozygosity, relatedness (Allendorf *et al.* 2010), these new genomic approaches provide an opportunity to tackle new questions regarding regions of the genome that underlie fitness-related traits (i.e., adaptive variation; Kohn *et al.* 2006; Harrisson *et al.* 2014; Mable 2018). While the promise of a conservation genomic approach has been heralded for well over a decade (Luikart *et al.* 2003), the uptake of these technologies by conservation management has been slow (Shafer *et al.* 2015; Galla *et al.* 2016).

This time lag between technology availability and implementation (also termed the ‘conservation genomics gap’; Shafer *et al.* 2015) may be caused by several interconnected issues, including a disconnect between conservation genetic researchers and practitioners (Knight *et al.* 2008; Taylor *et al.* 2017), the time it takes for geneticists to upskill in bioinformatic expertise (McCormack *et al.* 2013; Shafer *et al.* 2015; Galla *et al.* 2016), and initial expense for HTS sequence production and generation of genomic resources (e.g., a high-quality reference genome). With that said, sequencing costs are dropping precipitously (Hayden 2014) but see also (Muir *et al.* 2016) and affordable reduced representation genomic approaches provide the ability to produce high-density marker sets, even in the absence of a reference genome (i.e., de novo marker discovery (Narum *et al.* 2013). While it is possible to discover SNPs de novo, reference-guided approaches to SNP discovery offer many advantages, including enhanced computational efficiency, improved accuracy at low sequencing depth, higher confidence in identifying sequence contamination, greater ability

to identify the location of SNPs, improved performance in determining linkage disequilibrium between SNPs, and greater ability to identify differences between paralogous and repetitive sequences from true SNP variants (Davey *et al.* 2011; Ilut *et al.* 2014; Andrews *et al.* 2016; Oyler-McCance *et al.* 2016). Reference genomes also allow for identifying variants in annotated gene regions, which is necessary for identifying adaptive variation (Andrews *et al.* 2016). While reference genomes are preferred for conservation genomic research, they are often unavailable for threatened species or out of reach for resource-constrained conservation projects (e.g., Waldron *et al.* 2013).

There has been an exponential increase in the number of available eukaryotic genomes for non-model species that may be used as a reference (Ellegren 2014), including the outputs from various genome consortiums (e.g., Genome 10K, Genome 10K Scientists 2009; Bird 10,000 Genomes Project (B10K), Zhang *et al.* 2014; 5000 Insect Genome Project (i5K), Robinson *et al.* 2011; 1000 Plants Project (1KP), Matasci *et al.* 2014; Oz Mammalian Genomics, Duchêne *et al.* 2017; Earth BioGenome Project, Lewin *et al.* 2018). Readily available conspecific reference genomes for threatened species will likely enable faster uptake of a conservation genomics approach, for example, by avoiding the time and expenditure of sequencing and assembling a high-quality genome *de novo*. However, in many instances, the next best available resource may be a genome from a closely related species. There has been discussion on the utility of closely related reference genomes for reference-guided genome assembly (i.e., Card *et al.* 2014; Lischer & Shimizu 2017).

Additionally, there are many research studies to date that have used closely related reference genomes for SNP discovery using reduced-representation and whole genome resequencing (hereafter, resequencing) approaches (e.g., Der Sarkissian *et al.* 2015; Nuijten *et al.* 2016; Ng *et al.* 2017; Westbury *et al.* 2018).

Birds offer an exceptional opportunity to study the utility of SNP discovery using closely related reference genomes to inform conservation management. In comparison with other vertebrates, bird genomes are relatively small (~0.93–1.3 Gb), compact (i.e., low repetitive elements), and conserved between species (Organ *et al.* 2007; Zhang *et al.* 2014). Also, the availability of bird reference genomes has increased, due in part to the efforts of individual research groups that produce genomes to answer questions regarding primary production (e.g., chicken, *Gallus gallus*, Consortium 2004; the turkey, *Meleagris gallopavo*, Dalloul *et al.* 2010), evolution (e.g., zebra finch, *Taeniopygia guttata*, Warren *et al.* 2010; Galapagos cormorant, *Phalacrocorax harrisi*, Burga *et al.* 2017), and conservation (e.g., ‘amakihi/Hawaiian honeycreeper, *Hemignathus virens*, Callicrate *et al.* 2014; ‘alalā/Hawaiian crow, *Corvus hawaiiensis*, Sutton *et al.* 2018; kākāpō, *Strigops habroptilus*, Vertebrate Genomes Project 2018; kakī/black stilt, *Himantopus novaezelandiae*, this study). A substantial increase in the number of reference genomes available for birds can also be attributed to the efforts of B10K (Zhang *et al.* 2014; Peona *et al.* 2018), the international consortium whose goal is to produce a genome for every known species of bird. To date, B10K has published 38 de novo bird reference genomes (Zhang *et al.* 2014). These genomes, along with others that were available at the time of publication, make genomic resources available for at least one individual in almost every order of class Aves (Zhang 2015). The next phase of B10K will include genomes representing one species from every bird family ( $n = 240$ , Zhang 2015), increasing the availability of conspecific or closely related reference genomes for conservation research.

Here, we explore the utility of closely related reference genomes for SNP discovery using a critically endangered wading bird, the kakī, as proof of concept. Once found on the North and South Islands of New Zealand, kakī experienced significant population decline



throughout the 20th century due to habitat loss and degradation, and the introduction of mammalian predators. Today, there are approximately 132 kakī remaining (New Zealand Department of Conservation, *unpublished data*) and the population is contingent upon intensive management (Reed 1998; Sanders & Maloney 2002), including a captive breeding and rearing programme that uses genetic-based estimates of relatedness to pair distantly related individuals in captivity (Hagen *et al.* 2011). Beyond kakī, many programmes for threatened species incorporate neutral genetic measures (e.g., nucleotide diversity, individual heterozygosity or inbreeding, and relatedness) into management plans to minimise inbreeding (Ford *et al.* 2018) and loss of diversity (Pacioni *et al.* 2015; Jordan *et al.* 2016) to reduce extinction risk (Spielman *et al.* 2004; O’Grady *et al.* 2006).

To demonstrate that closely related reference genomes can yield sufficient SNPs to estimate diversity measures in threatened species, we map kakī genotyping-by-sequencing (GBS) and resequencing reads to genomes from members across the order Charadriiformes, including a conspecific reference genome (kakī, family: Recurvirostridae, *H. novaeseelandiae*), and members of the same genus (pied stilt, family: Recurvirostridae, *H. himantopus*), family (pied avocet, family: Recurvirostridae, *Recurvirostra avosetta*), and order (killdeer, family: Charadriidae, *Charadrius vociferus*) (Figure 4.1). Members from this comparison represent a wide evolutionary time scale: estimates based on traditional single-locus phylogenetic approaches suggest Charadriidae and Recurvirostridae diverged approximately 69 million years ago, avocets (genus: *Recurvirostra*) and stilts (genus: *Himantopus*) diverged approximately 36.9 million years ago, and kakī and pied stilt diverged approximately 1 million years ago (Wallis 1999; Baker *et al.* 2007; but see Jarvis *et al.* 2014) (Figure 4.1). SNPs discovered from these reference-guided assemblies were then compared

using estimates of diversity relevant to the conservation management of threatened species, including nucleotide diversity, individual heterozygosity, and relatedness.

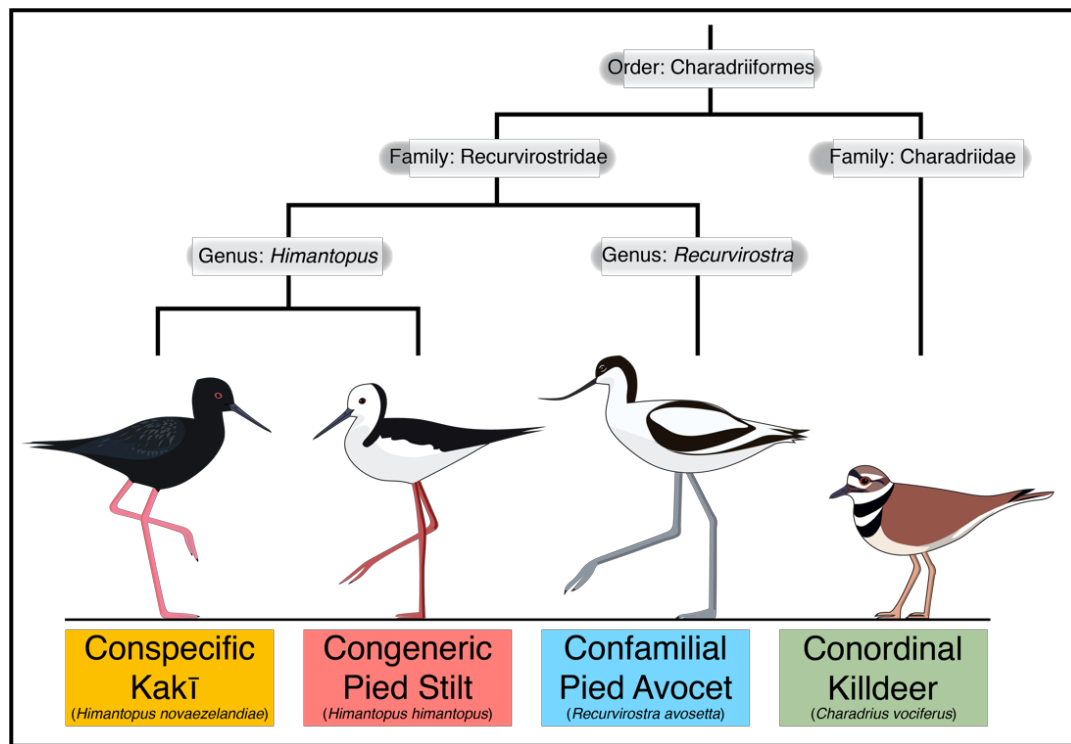


Figure 4. 1: Evolutionary relationships between species with reference genomes used in this proof of concept. The evolutionary tree indicates topology between taxa within the order Charadriiformes. Evolutionary tree is not to scale.

### 4.2.3: Materials and Methods

#### 4.2.3.1: Tissue Sampling and DNA Extractions

Kakī blood samples were collected during routine health checks by the New Zealand Department of Conservation (DOC) at the captive breeding facilities in Twizel (DOC) and Christchurch (Isaac Conservation and Wildlife Trust), New Zealand, by approval of the DOC Animal Ethics Committee (AEC #283). These samples were stored in 95% molecular grade ethanol at  $-80^{\circ}\text{C}$  at the University of Canterbury. Pied stilt blood samples were collected from one female and one male during routine health checks at Adelaide Zoo, with samples provided under the Royal Zoological Society of South Australia Specimen Licence Agreement (Import Permit: 2016061954). Pied stilt samples were stored in EDTA at  $-20^{\circ}\text{C}$  at the

University of Otago. The pied avocet blood sample was collected from a single individual from Hamburger Hallig, Germany, under a permit from the Ministry of Energy, Agriculture, the Environment, Nature and Digitization of the federal state of Schleswig-Holstein, Germany (Permit: V312-7224.121-37 [42-3/13]). Pied avocet samples were stored on filter paper at  $-20^{\circ}\text{C}$  at the University of Kiel.

Genomic DNA for kakī and pied stilt reference genomes was extracted at the University of Otago using a Thermo Scientific™ MagJET™ Genomic DNA Kit (Waltham, USA) following manufacturer specifications. DNA was isolated for the pied avocet sample at the University of Kiel Institute for Clinical Molecular Biology (hereafter, IKMB) by adding 400  $\mu\text{L}$  of phosphorus buffered saline solution (PBS) to dried blood and using the Qiagen® QIAmp® DNA Blood Mini QIAcube® Kit (Hilden, Germany) following the manufacturer specifications. Genomic DNA for the kakī genotyping-by-sequencing (GBS) and resequencing approaches was extracted at the University of Canterbury using a lithium chloride chloroform extraction method (see Supplement 1 for details). Genomic DNA for all extractions were analysed for quality using a NanoDrop™ Spectrophotometer and for quantity using an Invitrogen™ Qubit™ Fluorometer.

#### **4.2.3.2: Reference Genome Library Preparation and Sequencing**

Paired-end libraries for the kakī and pied stilt were prepared at the University of Otago using the Illumina TruSeq® DNA PCR-free protocol according to manufacturer specifications, with genomic DNA fragmented to 350 bp. End repair and adapter ligation for sequence barcoding were carried out and libraries were indexed with unique 6 bp sequences. Sequencing of kakī and pied stilt libraries was completed by New Zealand Genomics Limited (NZGL), where sample libraries were pooled with three additional stilt

samples and spread across five lanes of a flow cell for 2 × 125 bp sequencing on an Illumina HiSeq 2500.

Paired-end libraries for the pied avocet were prepared using the TruSeq<sup>®</sup> DNA Nano Library Prep protocol according to manufacturer specifications, with genomic DNA fragmented to 350 bp. Library preparation and sequencing for the pied avocet was completed at IKMB using one lane of a flow cell on an Illumina HiSeq 4000 for 2 × 150 bp sequencing.

#### **4.2.3.3: Reference Genome Sequence Processing and Assembly**

##### *4.2.3.3.1: Kakī and Australian Pied Stilt*

Raw kakī and pied stilt sequence reads were evaluated for quality using FastQC v. 0.11.5 (Andrews 2010). To test for exogenous contamination, the presence and abundance of non-avian reads was estimated by randomly subsampling 5000 reads from each library and searching these reads against the NCBI nucleotide database using BLAST (Altschul *et al.* 1990).

Illumina adapters used for sequence barcoding were removed using Trimmomatic v. 0.35 (Bolger *et al.* 2014). Low quality bases were trimmed using ConDeTri v. 2.3 (Smeds & Künstner 2011) with default settings. Read deduplication was carried out with ConDeTri, using the first 50 bp of both reads in a pair for comparisons. Raw reads were analysed using SGA-preqc v. 0.9.4 (Simpson 2014) to generate estimates of genome size and heterozygosity. To determine the level of expected heterozygosity in the genome and assess potential signatures of contamination, paired-end reads were analysed using KmerGenie (Chikhi & Medvedev 2013).

Trimmed sequences were assembled with SOAPdenovo2 (Luo *et al.* 2012) following initial testing of several assemblers and varying k-mer values. Draft assembly metrics were

independently assessed with the assembly metrics script generated for Assemblathon (Bradnam *et al.* 2013). BUSCO v. 3.0.1 (Simão *et al.* 2015; Waterhouse *et al.* 2017) was used to determine completeness of the assembly outputs based on expected gene content using an avian ortholog set derived from OrthoDB v. 9 (Zdobnov *et al.* 2016) and the chicken as reference. Both assembly metrics and BUSCO scores were used to determine the highest quality assemblies.

Trimmed sequence reads were used to close gaps between scaffolds in the highest quality assemblies for kakī and pied stilt with GapCloser v. 1.12 (Luo *et al.* 2012). Scaffolds shorter than 5 kbp were removed, and genomes were syntenically aligned against the chicken reference genome (version 5.0, GenBank Assembly GCF\_000002315.5) using Chromosome in Satsuma v. 3.1.0 (Grabherr *et al.* 2011) to generate pseudochromosome-level assemblies by aligning the draft assembly scaffolds against the chicken genome, and retaining orthologous regions. Final drafts of kakī and pied stilt genomes are available (see Data Availability section).

#### 4.2.3.3.2: *Pied Avocet*

Raw pied avocet sequence reads were evaluated for quality using FastQC v. 0.11.5 (Andrews 2010). To remove low quality reads, paired-end data was trimmed for Illumina adapter contamination and low quality bases using Skewer v. 0.2.2 (Jiang *et al.* 2014) with a mean Phred-score of 20, end-trim quality of 30, and a minimum length of 54 bp. Raw reads were analysed with SGA Preqc 0.10.15 (Simpson 2014) and KmerGenie (Chikhi & Medvedev 2013) to estimate heterozygosity and potential signatures of contamination. These analyses indicated high expected heterozygosity (0.3%) compared to other birds. To eliminate highly abundant repeats and sequencing errors, a digital normalisation was conducted using Khmer 2.1.1 (Crusoe *et al.* 2015).

Pied avocet trimmed sequences were assembled using Velvet 1.2.10 (Zerbino & Birney 2008) following initial testing with Meraculous-2D v. 2.2.5.1 (Goltsman *et al.* 2017), which failed to produce a high-quality assembly due to an overabundance of incorrectly merged diplotigs (i.e., contig pairs that share a unique k-mer at both ends (Chapman *et al.* 2016)). To evaluate the misassemblies, a second assembly was done with Velvet using default parameters. All contigs were aligned against the assembly using LAST (Kielbasa *et al.* 2011), with the -uNEAR seeding parameter. Alignments were filtered for trivial self-vs-self perfect alignments, with only single high-scoring pairs per sequence over 99% identical kept. These alignments revealed an unusual number of large and frequent indels (> 3 bp, higher than the default Velvet parameter for allowed gaps in graph bubbles) in extremely similar contigs, and therefore a final Velvet assembly was run with adjusted parameters (-ins\_length 410, -max\_branch\_length 50, -max\_divergence 0.1, -max\_gap\_count 10).

Assembled scaffolds were analysed with GapCloser v. 1.12 (Luo *et al.* 2012) to decrease gaps in the assembly. The gap-closed assembly was then aligned against the chicken genome using LAST (Kielbasa *et al.* 2011) and the Chromosomer (Tamazian *et al.* 2016) toolkit was used to construct superscaffolds. The final draft of the pied avocet genome is available (see Data Availability section).

#### 4.2.3.3.3: Killdeer

A killdeer genome was published in the ordinal phase of the B10K project (Zhang *et al.* 2014). To improve the assembly, a full de novo approach was used to construct a low-level base-accurate assembly. The data used in the original assembly of killdeer was downloaded from the GigaDB website (GigaDB 2014). This consisted of 12 libraries of Illumina sequence data, including five paired-end libraries with insert sizes ranging from 170 bp to 800 bp and seven mate-pair libraries of insert sizes ranging from 2000 bp to 20,000 bp.

FastQC v. 0.11.5 (Andrews 2010) was used to evaluate the quality of the Illumina data, as well as assess the contamination levels present in the samples. All paired-end libraries consisted of paired 100 bp reads, whereas mate-pair libraries were constructed of paired 50 bp reads. There was no evidence of any significant DNA contamination, but the per-base Phred-scores showed a consistently lower quality early in the reads. Due to the issues observed in the FastQC reports, reads were trimmed using Skewer v 0.2.2 (Jiang *et al.* 2014) to a minimum Phred-score of 30 and any read pair where at least one of the mates was trimmed to a length of < 32 bp was discarded.

Trimmed sequences were assembled using AllPaths-LG (Gnerre *et al.* 2011; Ribeiro *et al.* 2012) following initial testing of several assemblers and varying k-mer values. The first run was made with the two 170 bp libraries and the complete collection of mate-pair libraries. As part of the AllPaths-LG pipeline, a set of diagnostic data was generated, including estimates of genome size, error rates, and SNP rates. Three of the mate-pair libraries were removed from subsequent analysis after low levels of utilisation were detected due to failed library construction.

The new draft assembly was aligned against the original killdeer reference genome produced by Zhang *et al.* (2014) using the program LAST (Kielbasa *et al.* 2011), which identified areas of conflict between the original and new draft killdeer genomes (e.g., short gaps, abundance of small indels, and poor resolution in heterozygous regions in the original genome). A custom set of scripts, 'SemHelpers' (Moraga 2017b), was written to consolidate the changes detected via the genome-wide alignments into the existing reference genome. The resulting assembly has almost identical metrics when compared to the original assembly (Zhang *et al.* 2014), given the method used. Post-correction alignments between the final assembly and the original reference genome show identities between 98 and 99%.

Quality of all final draft assemblies was assessed with the Assemblathon metrics script (Bradnam *et al.* 2013) and completeness assessed with BUSCO v. 3.0.1 (Simão *et al.* 2015; Waterhouse *et al.* 2017) using the avian ortholog set and the chicken as reference. The final draft of the killdeer genome is available (see Data Availability section).

#### **4.2.3.4: Genotyping-by-Sequencing**

Genotyping-by-sequencing (GBS), a reduced-representation genomics approach, was used to produce genome-wide SNPs for kakī. Briefly, GBS reduces genome complexity by sequencing regions that flank restriction enzyme cut sites (Elshire *et al.* 2011). The GBS data presented here were produced following the Elshire *et al.* (2011) method, using 50 ng of genomic DNA with 0.72 ng of total adapters and the restriction enzyme ApeKI.

Because the kakī samples were collected during two different breeding seasons, library preparation and sequencing were completed in two separate batches. The first batch included captive parents and offspring from the 2015-2016 breeding season and other individuals sampled from 2014–2015 that represent diverse lineages based on the kakī pedigree ( $n = 52$ ; pedigree data not shown). This batch was sequenced with paired-end,  $2 \times 100$  bp reads on one lane of an Illumina HiSeq 2500 through NZGL. The second batch consisted of captive parents and offspring from the 2016-2017 breeding season plus one wild individual sampled in 2014 who represented a diverse lineage based on the pedigree ( $n = 47$ ). This batch was sequenced with paired-end,  $2 \times 150$  bp reads on one lane of an Illumina X Ten through CustomScience, Ltd. To assess the impact of batch effects (i.e., library and lane biases, Leigh *et al.* 2018), 10 individuals were represented in both batches to ensure similar genetic distance estimates were produced by each duplicated sample independently (see Table Appendix E.1 for individual sample sequencing details).



FastQC v. 0.11.4 (Andrews 2010) was used to evaluate the quality of the raw Illumina data, as well as assess the contamination levels present in the samples. Paired-end reads were demultiplexed and barcodes were trimmed using Axe (Murray & Borevitz 2017) with a maximum mismatch of 1. To minimise batch effects (Leigh *et al.* 2018) and address sequence quality, reads from the 2016-2017 breeding season were trimmed to a maximum length of 100 bp using Skewer (Jiang *et al.* 2014). To remove low quality data, reads were filtered to discard short reads (< 32 bp) and reads with mean quality scores less than 30.

In order to be read by downstream pipelines, new single-end barcodes were generated for the ApeKI enzyme using the programme GBSX (Herten *et al.* 2015) and appended to the forward-end of reads through a custom Perl script, 'mux\_barcodes' (Moraga 2017a). For this study, the Tassel 5.0 (Glaubitz *et al.* 2014) pipeline was used for SNP discovery and genotyping with GBS data. Due to the double-barcoding scheme of the GBS data generated here, a new class of enzymes was created specifically for Tassel 5 to add the enzyme cut site remnant, together with the reverse barcodes, as recognition sites for these datasets. The Tassel 5.0 GBSv2 pipeline was used with tag database and export plugins specifying a k-mer length of 64, a minimum k-mer length of 20, a minimum Phred-score of 30, and a minimum tag count of 10. Bowtie2 (Langmead & Salzberg 2012) was used to align tags to the each draft reference genome using the *--very-sensitive* presetting. The Tassel 5.0 GBSv2 discovery SNP caller plugin was run with a minimum minor allele frequency (-mnMAF) of 0.05 and a minimum locus coverage (-mnLCov) of 0.75. VCFtools v. 1.9 (Danecek *et al.* 2011) was used to filter the dataset to a set of bi-allelic SNPs, with an average minimum SNP depth of 5, and 90% of all SNPs being shared amongst individuals. To minimise statistical bias of linkage disequilibrium, the data set was pruned for linkage disequilibrium using BCFtools v. 1.9 (Li *et al.* 2009) with  $r^2$  set to 0.8 and a window size of

1000 sites. To ensure a more even spread of SNPs throughout the genome, VCFTools v. 1.9 (Danecek *et al.* 2011) was used to reduce the number of SNPs to 1 SNP within 64 bp, which is the designated size of a GBS tag using Tassel 5.0. VCFs of the filtered data set are available (see Data Availability section).

In order to evaluate whether the same SNPs were likely to be mapped using different reference genomes, a custom script, 'pancompare' (Moraga 2018), was used to compare pairs of tags in SAM files that are unique or shared between Tassel 5.0 runs using different reference genomes. This method uses tag pair mapping as a proxy for SNP discovery, under the assumptions that tags all start at the restriction cut site and intersecting pairs of tags are likely to discover the same SNPs using different reference genomes.

#### **4.2.3.5: Resequencing**

In addition to a reduced representation approach, we also resequenced kakī genomes from 36 individuals for SNP discovery and genotyping. These individuals include parents and offspring from the 2015-2016 and 2016-2017 breeding seasons ( $n = 24$ ) and other individuals sampled between 2014-2017 that represent diverse lineages based on the pedigree ( $n = 12$ ). Libraries were prepared at IKMB using a TruSeq® Nano DNA Library Prep kit following the manufacturer's specifications. Libraries were sequenced across 34 lanes on a HiSeq 4000 at the IKMB.

FastQC v. 0.11.4 (Andrews 2010) was used to initially evaluate the quality of the raw Illumina data, as well as assess the contamination levels present in the samples. Reads were trimmed for the Illumina barcode and for a Phred-score of 20 using Trimmomatic (Bolger *et al.* 2014). Reads were mapped to each indexed genome using Bowtie2 (Langmead & Salzberg 2012) with the *--very-sensitive* presetting. Resulting SAM files were converted to BAM files and read coverage was analysed using mpileup with Samtools v. 1.9 (Li *et al.*

2009). To improve the computational efficiency of mpileup, a custom Perl script ‘split\_bamfile\_tasks.pl’ (Moraga 2018) was created to subdivide BAM files and run them in parallel. SNPs were detected, filtered, and reported using BCFtools v.1.9 (Li *et al.* 2009). Filtering settings included biallelic SNPs with a minor allele frequency >0.05, an average mean depth >10, and a Phred-score >20. BCFtools was used to filter for a maximum of 10% missing data per site. Resulting SNPs were pruned for linkage disequilibrium using BCFtools with  $r^2$  set to 0.8 and a window size of 1000 sites. To ensure a more even spread of SNPs throughout the genome, VCFtools v. 1.9 (Danecek *et al.* 2011) was used to reduce the number of SNPs to 1 marker within 150 bp, which is the length of resequencing reads. VCFs of the filtered data set are available (see Data Availability section).

#### 4.2.3.6: Diversity Estimates

Nucleotide diversity ( $\pi$ ) and individual heterozygosity ( $H_S$ ) were estimated using VCFTools v. 1.9 (Danecek *et al.* 2011). Pairwise relatedness ( $R$ ) matrices were produced using KGD (Dodds *et al.* 2015), a programme that estimates relatedness while taking into account read depth of HTS data. To ease downstream mantel tests, pairwise  $R$  values were scaled so that self-relatedness of all individuals was equal to 1 using the formula:

$$M_S = D \times M_O \times D$$

where  $M_S$  is the scaled matrix,  $M_O$  is the original matrix, and  $D$  is a diagonal matrix with elements:

$$D = 1/\sqrt{\text{diag}(M_O)}$$

To compare  $H_S$  estimates generated from different reference genomes using GBS and resequencing data, analysis of variance (ANOVA) and Tukey multiple comparisons of means tests were performed using a linear mixed effects model with lme4 (Bates *et al.* 2015) to account for repeated measures (i.e., repeated individuals mapped to all four reference

genomes). Mantel tests with 1000 iterations were used to test whether scaled pairwise  $R$  matrices using different reference genomes were significantly similar compared to a null distribution. Correlations were conducted between estimates of  $H_S$  and  $R$  (not including self-relatedness) using different reference genomes using Spearman's rank ( $r_s$ ), which accounts for the inherently non-normal distribution of the  $R$  estimates.

#### 4.2.4: Results

##### 4.2.4.1: Reference Genome Sequencing and Assembly

Library sequencing produced 226–307 million paired-end sequences for each kakī, pied stilt, and avocet sample. Average sequencing depth was 52× for kakī, 51× for pied stilt, and 70× for avocet, based on an expected genome size of 1.2 Gb. Genomes produced were between 1.02–1.22 Gb in total length (Table 4.1), which is within the expected range for avian genomes (Gregory 2001). Scaffold N50 sizes ranged from 3.66 to 105.71 Mb. The total number of scaffolds ranged from 67 to 15,167. BUSCO assessment indicated the presence of at least 82.4% of the orthologs from the avian database. Combined, these estimates indicate that the assembled genomes have high genome completeness.

Table 4. 1: Genome assembly metrics for the genomes assembled in this study.

Species	Total Assembly Length (Gb)	Total Scaffolds	Scaffold N50 (bp)	Longest Scaffold (bp)	Average Scaffold Length (bp)	Complete Single-Copy BUSCOs (%)
Kakī	1.18	523	105,710,992	238,324,410	2,254,638	91.0
Pied Stilt	1.12	1443	99,457,149	221,521,436	773,955	85.9
Avocet	1.02	67	87,059,367	184,945,080	15,204,176	82.4
Killdeer	1.22	15,167	3,657,525	21,923,840	80,436	92.5

##### 4.2.4.2: SNP discovery and diversity estimates — GBS

After demultiplexing and initial read filtering, kakī GBS sequencing resulted in a total of 802.4 million reads for 88 individuals (mean =  $9.1 \pm \text{S.D. } 4.9$  million reads per individual). Five of these individuals were subsequently removed from the study after SNP filtering for having low average sample depths across sites ( $<4\times$  depth using conspecific reference

genome). The resulting 82 individuals have an average depth of 11.71–18.51x, with average missingness of 2–4% depending on the reference genome used (Table 4.2).

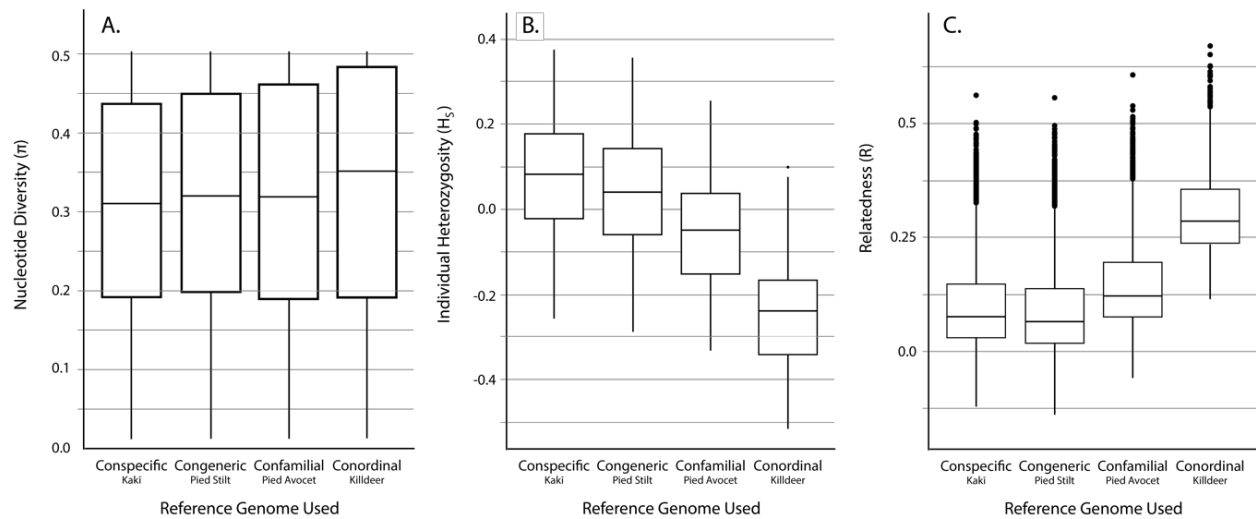
*Table 4. 2: Mapping statistics , single nucleotide polymorphisms (SNPs) discovered, SNP descriptive statistics, and average diversity statistics from genotyping-by-sequencing (GBS) reads mapped to different reference genomes.  $\pi$ : nucleotide diversity,  $H_s$ : individual heterozygosity,  $R$ : pairwise relatedness ( $\pm$  S.D. for each measure).*

Reference Genome	No. of Mapped Tag Pairs	% Tags Shared with Kakī Mapping	No. Unfiltered SNPs	No. Filtered SNPs	Average Missingness	Average Depth	Average $\pi$	Average $H_s$	Average $R$
Kaki	392,652	100	634,695	19,396	0.04 $\pm$ 0.04	13.73 $\pm$ 6.53	0.31 $\pm$ 0.14	0.07 $\pm$ 0.15	0.11 $\pm$ 0.12
Pied Stilt	372,906	91.04	604,573	18,625	0.04 $\pm$ 0.04	11.71 $\pm$ 5.52	0.32 $\pm$ 0.14	0.03 $\pm$ 0.15	0.10 $\pm$ 0.12
Avocet	316,978	83.10	481,532	18,398	0.03 $\pm$ 0.04	13.90 $\pm$ 6.58	0.31 $\pm$ 0.15	-0.06 $\pm$ 0.14	0.15 $\pm$ 0.11
Killdeer	151,546	72.42	242,493	10,440	0.02 $\pm$ 0.03	18.51 $\pm$ 8.77	0.33 $\pm$ 0.15	-0.25 $\pm$ 0.14	0.30 $\pm$ 0.09

The number of GBS tag pairs mapped to each reference genome was greatest using a conspecific reference genome, with fewer tag pairs mapped the more phylogenetically distant the reference genome became (Table 4.2). Results from our analysis with ‘pancompare’ (Moraga 2018) indicate that more tags from the congeneric mapping were shared with those mapped to a conspecific reference genome (91.04%) than more distantly related genomes (confamilial = 83.10% and conordinal = 72.42%; Table 4.2). Tag pairs always start at the GBS restriction enzyme cut site, making direct comparisons of tags mapped across different genomes possible. Because more mapped tags were shared between closely related genomes, these results suggest that SNPs discovered with the congeneric reference genome are more likely the same as those discovered with the conspecific reference genome than those discovered with the confamilial or conordinal references.

The number of unfiltered and filtered SNPs discovered was greatest when using a conspecific reference genome, with fewer SNPs discovered the more phylogenetically distant the reference genome became (Table 4.2). Despite the differences in number of

SNPs discovered with each reference genome, average nucleotide diversity ( $\pi$ ) was similar across datasets (average  $\pi = 0.31$ – $0.33$ , Table 4.2, Figure 4.2A).



**Figure 4. 2: Distribution of different diversity estimates using SNPs discovered with GBS reads mapped against different reference genomes. (A) Nucleotide diversity ( $\pi$ ), (B) individual heterozygosity ( $H_s$ ), and (C) pairwise relatedness ( $R$ ) not including self-relatedness.**

Average individual heterozygosity ( $H_s$ ) estimates differed depending on the reference genome used (Table 4.2, Figure 4.2B). Results show that using different reference genomes produced significantly different levels of  $H_s$  from one another (ANOVA,  $p < 0.001$ ; Tukey Contrasts,  $p < 0.001$ ). Using a congeneric reference genome resulted in  $H_s$  estimates that are on average 3.4% less than using a conspecific reference genome, with a confamilial being 12.9% less, and a conordinal being 31.6% less. Despite significant differences in  $H_s$  depending on the reference genome used, estimates of  $H_s$  using different reference genomes were significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.996$ ) being higher than the conspecific and confamilial approaches ( $r_s = 0.990$ ) and conspecific and conordinal approaches ( $r_s = 0.963$ ; Figure 4.3A–C).

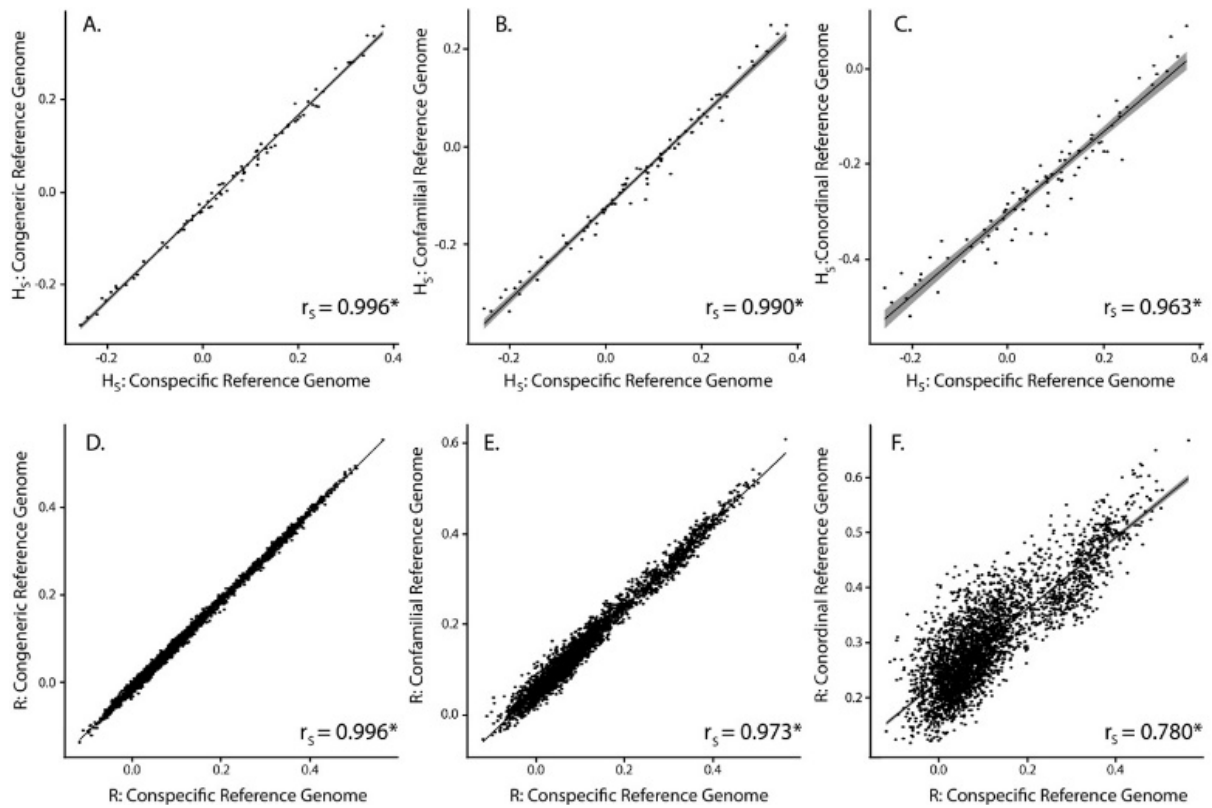


Figure 4. 3: Scatterplots showing individual point estimates of HS (A–C) and pairwise  $R$  estimates (D–F) using GBS reads mapped to different reference genomes. Self-relatedness estimates were not used in this analysis. Trend line in black, with 95% confidence intervals surrounding the trend line in gray. Spearman's correlation coefficient ( $r_s$ ) provided in the lower right corner of each scatterplot. \* indicates significance  $p < 0.001$ .

The range of scaled average pairwise estimates of relatedness ( $R$ ) shows a bimodal distribution, which reflects highly related individuals (siblings and parent-offspring relationships) along with more distantly related individuals that are captured in the study design. The range of scaled  $R$  values appeared different depending on the reference genome used, with average pairwise  $R$  in the conspecific and congeneric analyses being less than the confamilial and conordinal analyses (Table 4.2). Despite this pattern, the relationships between  $R$  using a conspecific reference genome and all other genomes were not significantly different (Mantel test,  $p < 0.001$ ). Estimates of pairwise  $R$  (not including self-relatedness) using different reference genomes were significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.996$ ) being higher than the conspecific and confamilial

approaches ( $r_s = 0.973$ ) and the conspecific and conordinal approaches ( $r_s = 0.780$ ; Figure 4.3D–F).

#### 4.2.4.3: SNP Discovery and Diversity Estimates — Resequencing

After demultiplexing and initial read filtering, the kakī resequencing resulted in a total of 4.8 billion reads for 36 individuals (mean =  $135.8 \pm 54.1$  million reads per individual).

After SNP filtering, these 36 individuals have an average depth of 13.95–17.44× with average missingness of 0.2% across all reference genomes used (Table 4.3).

*Table 4. 3: Alignment rates, single nucleotide polymorphisms (SNPs) discovered, SNP descriptive statistics, and average diversity statistics from resequencing reads mapped to different reference genomes.  $\pi$ : nucleotide diversity,  $H_s$ : individual heterozygosity,  $R$ : pairwise relatedness. ( $\pm$  S.D. for each measure).*

Reference Genome	Average Alignment Rate (%)	No. Unfiltered SNPs	No. Filtered SNPs	Average Missingness	Average Depth	Average $\pi$	Average $H_s$	Average $R$
Kaki	$94.6 \pm 0.50$	4,246,100	91,854	$0.002 \pm 0.005$	$17.44 \pm 6.79$	$0.35 \pm 0.13$	$-0.05 \pm 0.08$	$0.06 \pm 0.11$
Pied Stilt	$88.1 \pm 0.96$	8,438,866	89,419	$0.002 \pm 0.005$	$14.99 \pm 6.06$	$0.34 \pm 0.13$	$-0.05 \pm 0.08$	$0.06 \pm 0.11$
Avocet	$78.5 \pm 0.46$	24,333,620	143,343	$0.002 \pm 0.004$	$16.02 \pm 6.43$	$0.33 \pm 0.14$	$-0.05 \pm 0.07$	$0.11 \pm 0.11$
Killdeer	$64.8 \pm 4.89$	62,888,931	89,145	$0.002 \pm 0.004$	$13.95 \pm 5.54$	$0.32 \pm 0.13$	$0.25 \pm 0.07$	$0.03 \pm 0.13$

Average read alignment rates using Bowtie2 were highest when using a conspecific reference genome (94.6%), with fewer reads aligning with congeneric (88.1%), confamilial (78.5%), and conordinal reference genomes (64.8%, Table 4.3). In contrast to GBS, the number of unfiltered SNPs increased with phylogenetic distance of the reference genome, which is expected given resequencing SNPs are called by differences between reads and the reference. The number of SNPs discovered post filtering did not correspond with phylogenetic distance of the reference used, with the fewest filtered SNPs being discovered with the conordinal reference genome (89,145) and the most being discovered with the confamilial reference genome (143,343, Table 4.3). Similar to the GBS dataset, average  $\pi$  was similar across datasets generated using different reference genomes (average  $\pi = 0.32$ – $0.35$ , Table 4.3, Figure 4.4A).



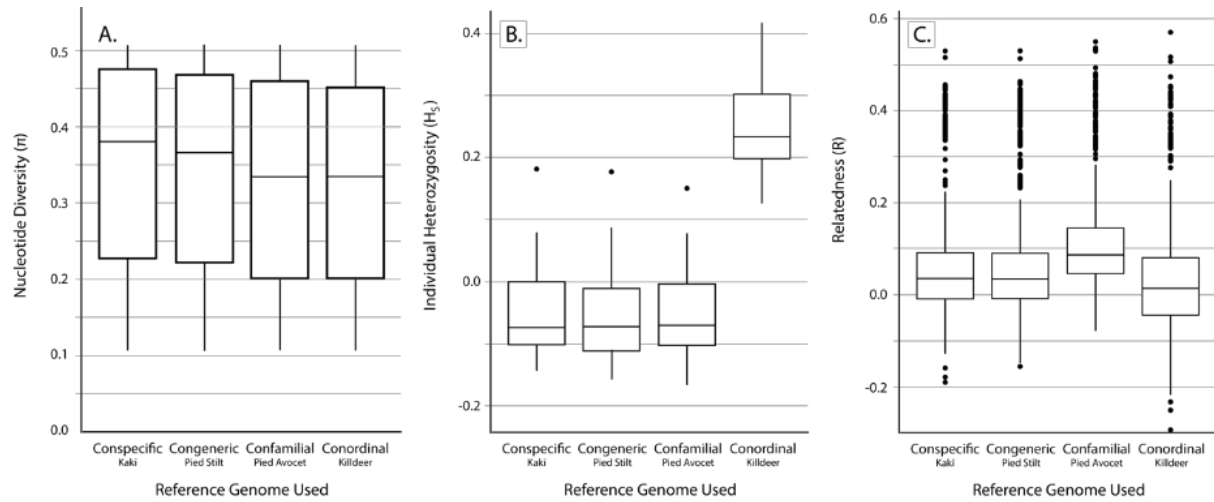


Figure 4. 4: Distribution of different diversity estimates using SNPs discovered with resequencing reads mapped against different reference genomes. (A) Nucleotide diversity ( $\pi$ ), (B) individual heterozygosity ( $H_s$ ), and (C) pairwise relatedness ( $R$ ). Self-relatedness estimates were not used in this analysis.

Results show that using a conordinal reference genome produced significantly higher levels of  $H_s$  than the conspecific, congeneric, or confamilial approaches (ANOVA,  $p < 0.001$ ; Tukey contrasts,  $p < 0.001$ ; Table 4.2, Figure 4.4B). Using a congeneric reference genome resulted in  $H_s$  estimates that are on average 0.40% less than using a conspecific reference genome, with a confamilial being 0.31% less, and a conordinal being 29.9% greater. Despite significant differences in  $H_s$  depending on the reference genome used,  $H_s$  using different reference genomes is significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.987$ ) being higher than congeneric and confamilial approaches ( $r_s = 0.981$ ) and congeneric and conordinal approaches ( $r_s = 0.823$ ; Figure 4.5A–C).

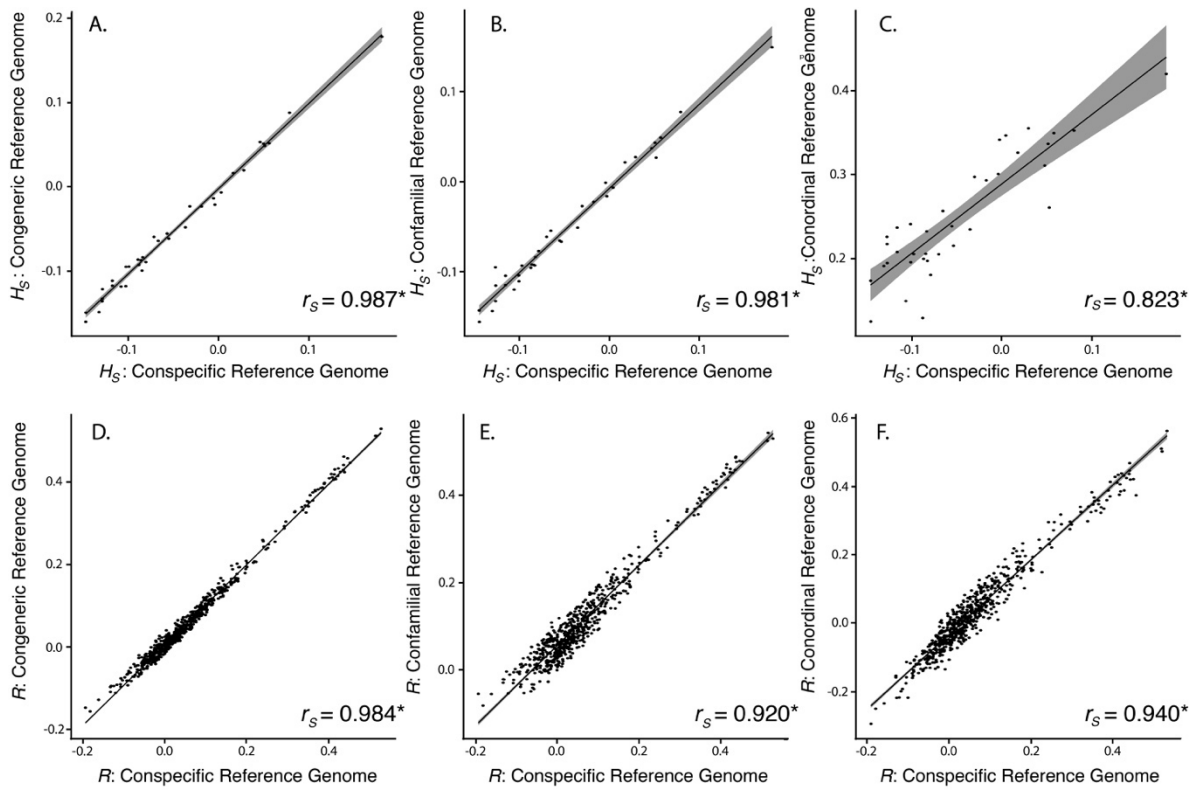


Figure 4.5: Scatterplots showing individual point estimates of HS (A–C) and pairwise R estimates (D–F) using resequencing reads mapped to different reference genomes. Self-relatedness estimates were not used in this analysis. Trend line in black, with 95% confidence intervals surrounding the trend line in gray. Spearman's correlation coefficient ( $r_s$ ) provided in the lower right corner of each scatterplot. \* indicates significance  $p < 0.001$ .

Similar to the GBS approach, the range of scaled average pairwise estimates of relatedness ( $R$ ) based on resequencing also shows a bimodal distribution, which reflects the relationships of individuals captured in the study design. Average scaled pairwise estimates of  $R$  were similar across all reference genomes used (Table 4.2, Figure 4.4C). The relationship between  $R$  using a conspecific reference genome and all other genomes were not statistically different compared to the null distribution (Mantel test,  $p < 0.001$ ). Scaled pairwise  $R$  (not including self-relatedness) using different reference genomes is significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.984$ ) being higher than conspecific and confamilial approaches ( $r_s = 0.920$ ) and conspecific and conordinal approaches ( $r_s = 0.940$ ; Figure 4.5D–F).

#### 4.2.5: Discussion

For species of conservation concern, limited conspecific genomic resources often impede inclusion of genomic data in conservation management strategies. Our proof of concept demonstrates that SNPs discovered using congeneric, confamilial, and even conordinal approaches yield diversity estimates that significantly correlate with estimates derived from SNPs discovered using a conspecific approach. Prior to this study, there was only one genome publicly available for the order Charadriiformes (i.e., the killdeer, Zhang *et al.* 2014). This study provides three additional high-quality de novo genome assemblies, all of which have practical applications for conservation.

The number of GBS tag pairs that aligned to each reference genome decreased the more phylogenetically distant the reference genome became. Because Tassel 5.0 calls SNPs based on differences among tag pairs (Glaubitz *et al.* 2014)—as opposed to differences between tag pairs and the reference genome—the number of unfiltered SNPs discovered also decreased the more phylogenetically distant the reference genome became. The same pattern was observed for the number of filtered SNPs. The ‘pancompare’ analysis of GBS tag data suggests that SNP discovery using the conspecific and congeneric reference genomes are more likely to yield the same markers compared to SNPs discovered using the confamilial or conordinal reference genomes.

The number of resequencing reads that aligned to each reference genome also decreased the more phylogenetically distant the reference genome became. Unlike GBS, the number of unfiltered SNPs increased with phylogenetic distance. This is to be expected because the resequencing discovery pipeline calls SNPs based on differences between reads and the reference genome (Li *et al.* 2009). The number of SNPs discovered post-filtering was

unexpected, however, as a similar number of SNPs were found in all but the confamilial reference approach, which resulted in  $\sim 1.5\times$  more SNPs than other reference-guided approaches. While the pied avocet genome shows signs of high completeness, complexities in the genome assembly due to high heterozygosity (Goltsman *et al.* 2017; Tigano *et al.* 2018) may have resulted in less complete regions leading to higher false discovery rates (Peona *et al.* 2018).

Using GBS and resequencing data, the average and range of  $\pi$  estimates did not differ greatly based on reference genome used. Larger differences between reference genomes used were observed when estimating  $H_S$ . Using GBS data, mean estimates of  $H_S$  decreased significantly the more distant the reference genomes became, with the use of a conordinal reference genome producing a marked decrease in  $H_S$  estimates compared to the use of a conspecific reference. This decrease in  $H_S$  corresponds to an increase in  $R$ , although not significantly so. These combined results are consistent with expectations because SNPs called by Tassel 5.0 are based on identifying mapped tag pairs (Glaubitz *et al.* 2014); the more phylogenetically distant a reference genome is, the more conserved a region has to be to successfully map a pair of tags. Therefore, with GBS we expect  $H_S$  to be lower and  $R$  to be higher the more phylogenetically distant the reference used is, given that variants at these conserved regions are less frequent.

Using resequencing data, conspecific, congeneric, and confamilial approaches produced  $H_S$  that were not significantly different from one another, with the only significant difference seen with the conordinal approach, which resulted in a significant increase in  $H_S$  compared to other reference genomes. Unlike GBS tags, there is not an immediate explanation for this pattern. However, it may be attributed to the fact that resequencing

reads, which are longer and are more representative of the whole genome, can be mapped to more divergent regions than GBS tags.

While the range of  $H_S$  and pairwise  $R$  values may be different depending on the reference genome used, all estimates produced using different reference genomes correlate significantly with one another. Our results suggest that using a more closely related reference genome (e.g., congeneric) over a more distant reference genome (e.g., conordinal) will yield SNPs that have higher correlation coefficients with estimates generated using a conspecific, and therefore, are likely to result in similar conservation recommendations. Ongoing work incorporating genomic based estimates of relatedness into software that informs captive pairing recommendations (e.g., PMx; Lacy *et al.* 2012) will indicate whether more distantly related reference genomes indeed produce statistically similar pairing recommendations, as our correlation results suggest. In the meantime, we anticipate even small changes in  $H_S$  and pairwise  $R$  estimates will not greatly affect conservation recommendations, as diversity estimates are often used in relative terms. For example, pairing recommendations for intensively managed populations that lack reliable pedigrees are routinely informed by genetic- or genomic-based pairwise estimates of relatedness (e.g., Hagen *et al.* 2011; Putnam & Ivy 2013; Willoughby *et al.* 2015; Hammerly *et al.* 2016). In practice, pairing recommendations are made based on the relative ranking of these estimates and not the absolute values. Similarly, when investigating heterozygosity-fitness correlations (e.g., Szulkin *et al.* 2010), relative rankings of  $H_S$  among individuals are more informative than absolute values.

Still, there may be some instances where absolute diversity values may be of interest (e.g., parentage assignment, or management of individuals that exhibit  $H_S$  below a cutoff score; Sandoval-Castillo *et al.* 2017). SNPs derived using the conordinal reference genome

provide markedly different ranges of  $H_s$  and pairwise  $R$  estimates and often the lowest correlation coefficients compared to SNPs derived from the conspecific reference genome. For birds, we recommend a confamilial reference genome as the most distant reference genome conservation researchers consider using for diversity estimates. However, this approach should be evaluated for use in other questions, such as the characterisation of adaptive variation (Andrews *et al.* 2016; Mable 2018).

The number of de novo bird genomes available to be used as reference is due to increase, especially as the next phase of B10K seeks to publish representative genomes for every recognised family of birds (Zhang 2015). However, we recommend evaluating the quality of publicly available genomes prior to use, as lower quality genomes may produce lower SNP yield due to fewer alignable regions, or greater false discovery rate where there are assembly errors (Trapnell & Salzberg 2009). Here, we re-assembled the available killdeer reference genome for two reasons. First, the raw data available from the European Bioinformatics Institute European Nucleotide Archive (EBI ENA) showed poor sequencing quality and mapping of this raw data to the existing reference suggested inconsistencies where poor quality reads were more abundant. Second, mapping of the long-insert mate-pair data from the project showed little to no support for many of the scaffolding connections present in the published genome. Due to these factors, we reassembled the genome using much more stringent data curation and more cautious scaffolding. Given this, when using a genome “off the shelf”, we recommend careful assessment of the original genome publication, keeping in mind that genomes assembled from multiple libraries or data types, with greater depth of sequencing coverage, and a more complete and contiguous assembly, will be of higher quality (Peona *et al.* 2018). When genomes with similar phylogenetic relationships are available, comparisons of synteny (Grabherr *et al.*

2011) and completeness (Peona *et al.* 2018) against the most closely related model genome may help identify which genome is most appropriate to use. Ultimately, the best way to assess existing genomic resources is to download the raw reads and evaluate them using tools such as FastQC (Andrews 2010) and SGA pre-QC (Simpson 2014), as we have done with the killdeer genome. Raw read quality may have the largest impact on final assembly quality, and initial quality checks will allow identification of any potential anomalies or limitations of the raw data that may have presented challenges to assembly, such as high heterozygosity (Kajitani *et al.* 2014; Goltsman *et al.* 2017; Tigano *et al.* 2018). If the raw data is of high quality, but there are inconsistencies between original reported statistics and those derived from raw reads, it may be worth investing in re-assembly to produce a genome of higher quality with greater confidence.

Indeed, re-assembly remains a more cost-effective option than starting a genome sequencing project from scratch. By our current (2018) estimates based on single libraries with paired-end reads, the use of a closely related high quality readily-available reference genome may save a conservation genomic project a minimum of EUR 6,500 in library preparations, sequencing, computational power, and assembly time (Table Appendix E.2, although prices subject to rapid change given new sequencing technologies). Among the 383 species in the order Charadriiformes, 51 are threatened with extinction (IUCN 2018). The families Laridae (gulls, terns, and skimmers) and Scolopacidae (sandpipers) contain particularly high numbers of threatened species (14 and 13, respectively). Along with the genomes produced in this paper, there are now genomes available for four additional families within Charadriiformes (i.e., Alcidae, Tigano *et al.* 2018; Charadriidae, Zhang *et al.* 2014; Recurvirostridae, here; and Scolopacidae, Küpper *et al.* 2016). Genome sequencing and assembly of one member of the Laridae family could benefit all 14 threatened species

within this family, and combined with the existing genomes available as reference within Scolopacidae, could save conservation groups up to EUR 169,000 in sequencing and assembly costs. Using existing genomic resources will not only reduce these costs, but also the time needed to produce a high-quality reference genome, thereby allowing for a faster uptake of conservation genomics approaches to produce robust information for conservation management.

#### *4.2.6: Conclusions*

Many threatened species management programmes rely on measures of diversity, including nucleotide diversity, heterozygosity, and relatedness, in guiding management decisions (Frankham 2010; Willoughby *et al.* 2015). While these measures have historically been calculated using small numbers of genetic markers, genomic markers offer the opportunity for increased resolution (Allendorf *et al.* 2010; Taylor 2015; Galla *et al.* 2016) and hence improved decision-making. Here, we have demonstrated that in the absence of a conspecific reference genome to map genomic sequence reads to, the availability of high-quality reference genome for a closely related species can provide highly correlated estimates for nucleotide diversity, individual heterozygosity, and relatedness. We anticipate the use of readily available reference genomes may provide resource-constrained conservation projects a way to minimise these costs and make a faster transition to using genomic data to improve conservation outcomes for threatened species.

#### *4.2.7: Data Availability*

The pied stilt Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession RSEF00000000. The version described in this paper is version RSEF01000000. The pied avocet genome raw reads have been deposited in Genbank under project number PRJNA508299. The reassembled killdeer genome is available



at <http://www.uconsert.org/data/>. Kakī are taonga (treasured) to Māori (the indigenous people of Aotearoa New Zealand), and as such the genomes obtained from taonga species are taonga in their own right. Therefore, the genome for kakī and all VCFs for GBS and resequencing will be made available on recommendation of the iwi (tribes) that affiliate as kaitiaki (guardians) for kakī. A local genome browser is available to view the kakī genome and all VCFs presented here at <http://www.uconsert.org/data/>, along with details on how to request access.

#### *4.2.8: Supplemental Materials*

The following are available online at [www.mdpi.com/xxx/s1](http://www.mdpi.com/xxx/s1), Table Appendix E.1: Samples used in Genotyping-by-Sequencing and resequencing analyses; Table Appendix E.2: Cost associated with genome sequencing and alignment; lithium chloride extraction protocol.

#### *4.2.9: Funding*

This research was funded by the Brian Mason Scientific & Technical Trust (SJG, TES), the Mohua Charitable Trust (TES), New Zealand Ministry of Business, Innovation and Employment Endeavour Fund (TES, AWS), and the Royal Society of New Zealand Rutherford Discovery Fellowship (MK).

#### *4.2.10: Acknowledgements*

We are grateful for the support of Te Rūnanga o Ngāi Tahu, Te Rūnanga o Arowhenua, Te Rūnanga o Waihao and Te Rūnanga o Moeraki. We thank Elshire Group Ltd. for their support in generating GBS data. We thank Sabine Müller of the Research and Technology Centre West Coast, Kiel University, for kindly providing access to the pied avocet blood sample, and the Royal Zoological Society of South Australia (Adelaide Zoo) for providing the pied stilt blood sample. Finally, we thank all members of the Kakī Recovery Group, especially

Simone Cleland and Cody Thyne, and the Isaac Conservation and Wildlife Trust, especially Anne Richardson, for their assistance with sample collection and their ongoing support. We also thank three anonymous reviewers for their helpful comments on the manuscript.

#### 4.2.11: References

- Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Review Genetics*, **11**(10), 697.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403-410.
- Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Review Genetics*, **17**(2), 81.
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Retrieved from <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Baker AJ, Pereira SL, Paton TA (2007) Phylogenetic relationships and divergence times of Charadriiformes genera: multigene evidence for the Cretaceous origin of at least 14 clades of shorebirds. *Biology Letters*, **3**(2), 205-210.
- Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1-48.
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15), 2114-2120.
- Bradnam KR, Fass JN, Alexandrov A, *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, **2**(1), 10.
- Burga A, Wang W, Ben-David E, *et al.* (2017). A genetic signature of the evolution of loss of flight in the Galapagos cormorant. *Science*, **356**(6341), eaal3345.
- Callicrate T, Dikow R, Thomas JW, Mullikin JC, Jarvis ED, Fleischer RC (2014) Genomic resources for the endangered Hawaiian honeycreepers. *BMC Genomics*, **15**(1), 1098.
- Card DC, Schield DR, Reyes-Velasco J, *et al.* (2014) Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. *PLoS One*, **9**(9), e106649.
- Chapman JA, Ho IY, Goltsman E, Rokhsar DS (2016) Meraculous2: fast accurate short-read assembly of large polymorphic genomes. *ArXiv e-prints*, 1608.01031.

- Chikhi R, Medvedev P (2013) Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, **30(1)**, 31-37.
- International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432(7018)**, 695.
- Crusoe MR, Alameldin HF, Awad S, *et al.* (2015) The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research*, **4**.
- Dalloul RA, Long JA, Zimin AV, *et al.* (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biology*, **8(9)**, e1000475.
- Danecek P, Auton A, Abecasis G, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics*, **27(15)**, 2156-2158.
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12(7)**, 499.
- Der Sarkissian C, Ermini L, Schubert M, *et al.* (2015) Evolutionary genomics and conservation of the endangered Przewalski's horse. *Current Biology*, **25(19)**, 2577-2583.
- Dodds KG, McEwan JC, Brauning R, Anderson RM, Stijn TC, Kristjánsson T, Clarke SM (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, **16(1)**, 1047.
- Duchêne DA, Bragg JG, Duchêne S, *et al.* (2017). Analysis of phylogenomic tree space resolves relationships among marsupial Families. *Systematic Biology*, **67(3)**, 400-412.
- Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, **29(1)**, 51-63.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6(5)**, e19379.
- Ford MJ, Parsons K, Ward E, *et al.* (2018) Inbreeding in an endangered killer whale population. *Animal Conservation*, **21(5)**, 423-432.
- Frankham R (2010) Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, **143(9)**, 1919-1927.
- Galla SJ, Buckley TR, Elshire R, *et al.* (2016). Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances. *Molecular Ecology*, **25(21)**, 5267-5281.

- GigaDB. (2014). Genomic data of the Killdeer (*Charadrius vociferus*). Retrieved from <http://dx.doi.org/10.5524/101007>
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*, **9(2)**, e90346.
- Gnerre S, MacCallum I, Przybylski D, *et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Science USA*, **108(4)**, 1513-1518.
- Goltsman E, Ho I, Rokhsar D (2017) Meraculous-2D: Haplotype-sensitive Assembly of Highly Heterozygous genomes. *ArXiv e-prints*.
- Grabherr MG, Haas BJ, Yassour M, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29(7)**, 644-652
- Gregory TR (2001) Animal genome size database. Retrieved from <http://www.genomesize.com/>
- Hagen EN, Hale ML, Maloney RF, Steeves TE (2011) Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the black stilt *Himantopus novaezelandiae*. *Ibis*, **153(3)**, 556-561.
- Hammerly SC, de la Cerda DA, Bailey H, Johnson JA (2016) A pedigree gone bad: increased offspring survival after using DNA-based relatedness to minimize inbreeding in a captive population. *Animal Conservation*, **19(3)**, 296-303.
- Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014) Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, **7(9)**, 1008-1025.
- Hayden EC (2014) The \$1,000 genome. *Nature*, **507**, 294-295.
- Herten K, Hestand MS, Vermeesch JR, Van Houdt JK (2015) GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics*, **16(1)**, 73.
- Ilut DC, Nydam ML, Hare MP (2014) Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. *BioMed Research International*, 2014.
- IUCN. (2018). The IUCN Red List of Threatened Species. Retrieved from <http://www.iucnredlist.org>

- Jarvis ED, Mirarab S, Aberer AJ, *et al.* (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**(6215), 1320-1331.
- Jiang H, Lei R, Ding SW, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**(1), 182.
- Jordan S, Giersch JJ, Muhlfeld CC, Hotaling S, Fanning L, Tappenbeck TH, Luikart G (2016) Loss of genetic diversity and increased subdivision in an endemic alpine stonefly threatened by climate change. *PLoS One*, **11**(6), e0157386.
- Kajitani R, Toshimoto K, Noguchi H, *et al.* (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**(8), 1384-1395.
- Kielbasa SM, Wan R, Sato K, Horton P, Frith M (2011) Adaptive seeds tame genomic sequence comparison. *Genome Research*, **21**(3), 487-493.
- Knight AT, Cowling RM, Rouget M, Balmford A, Lombard AT, Campbell BM (2008) Knowing but not doing: selecting priority conservation areas and the research–implementation gap. *Conservation Biology*, **22**(3), 610-617.
- Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. *Trends in Ecology and Evolution*, **21**(11), 629-637.
- Küpper C, Stocks M, Risse JE, *et al.* (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, **48**(1), 79.
- Lacy RC, Ballou JD, Pollak JP (2012) PMx: software package for demographic and genetic analysis and management of pedigreed populations. *Methods in Ecology and Evolution*, **3**(2), 433-437.
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4), 357.
- Leigh D, Lischer H, Grossen C, Keller L (2018) Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. *Molecular Ecology Resources*, **18**(4), 778-788.
- Lewin HA, Robinson GE, Kress WJ, *et al.* (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Science USA*, **115**(17), 4325-4333.
- Li H, Handsaker B, Wysoker A, *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078-2079.
- Lischer HE, Shimizu KK (2017) Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, **18**(1), 474.

- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Review Genetics*, **4(12)**, 981.
- Luo R, Liu B, Xie Y, *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1(1)**, 18.
- Mable BK (2018) Conservation of adaptive potential and functional diversity: integrating old and new approaches. *Conservation Genetics*, **20(1)**, 89-100.
- Matasci N, Hung LH, Yan Z, *et al.* (2014) Data access for the 1,000 Plants (1KP) project. *GigaScience*, **3(1)**, 17.
- McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66(2)**, 526-538.
- Moraga R (2017a) Mux Barcodes [Custom Perl Script]. Retrieved from [https://github.com/sgalla32/mux\\_barcodes](https://github.com/sgalla32/mux_barcodes)
- Moraga R (2017b) SemHelpers [Custom Perl Script]. Retrieved from <https://github.com/Lanilen/SemHelpers>
- Moraga R (2018) Pancompare [Custom Perl Script]. <https://github.com/Lanilen/pancompare>.
- Muir P, Li S, Lou S, *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, **17(1)**, 53.
- Murray KD, Borevitz JO (2017) Axe: rapid, competitive sequence read demultiplexing using a trie. *Bioinformatics*, **34(22)**, 3924-3925.
- Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22(11)**, 2841-2847.
- Ng EY, Garg KM, Low GW, Chattopadhyay B, Oh RR, Lee JG, Rheindt FE (2017) Conservation genomics identifies impact of trade in a threatened songbird. *Biological Conservation*, **214**, 101-108.
- Nuijten RJ, Bosse M, Crooijmans RP, *et al.* (2016) The use of genomics in conservation management of the endangered visayan warty pig (*Sus cebifrons*). *International Journal of Genomics*, **2016**.

- O'Grady JJ, Brook BW, Reed DH, Ballou JD, Tonkyn DW, Frankham R (2006) Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. *Biological Conservation*, **133**(1), 42-51.
- Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. *Nature*, **446**(7132), 180.
- Oyler-McCance SJ, Oh KP, Langin KM, Aldridge CL (2016) A field ornithologist's guide to genomics: Practical considerations for ecology and conservation. *Auk*, **133**(4), 626-648.
- Pacioni C, Hunt H, Allentoft ME, *et al.* (2015) Genetic diversity loss in a biodiversity hotspot: ancient DNA quantifies genetic decline and former connectivity in a critically endangered marsupial. *Molecular Ecology*, **24**(23), 5813-5828.
- Peona V, Weissensteiner MH, Suh A (2018) How complete are “complete” genome assemblies?—An avian perspective. *Molecular Ecology Resources*, **18**(6), 1188-1195.
- Putnam AS, Ivy JA (2013) Kinship-based management strategies for captive breeding programs when pedigrees are unknown or uncertain. *Journal of Heredity*, **105**(3), 303-311.
- Reed CEM (1998) *Management plan for captive black stilts*. Wellington, New Zealand: Biodiversity Recovery Unit, Department of Conservation.
- Ribeiro F, Przybylski D, Yin S, *et al.* (2012) Finished bacterial genomes from shotgun sequence data. *Genome Research*, **22**(11), 2270-2277.
- Robinson GE, Hackett KJ, Purcell-Miramontes M, *et al.* (2011) Creating a buzz about insect genomes. *Science*, **331**(6023), 1386-1386.
- Sanders MD, Maloney RF (2002) Causes of mortality at nests of ground-nesting birds in the Upper Waitaki Basin, South Island, New Zealand: a 5-year video study. *Biological Conservation*, **106**(2), 225-236.
- Sandoval-Castillo J, Attard CRM, Marri S, Brauer CJ, Möller LM, Beheregaray LB (2017) swinger: a user-friendly computer program to establish captive breeding groups that minimize relatedness without pedigree information. *Molecular Ecology Resources*, **17**(2), 278-287.
- Genome 10K Community of Scientists (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, **100**(6), 659-674.
- Shafer AB, Wolf JB, Alves PC, *et al.* (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, **30**(2), 78-87.

- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31(19)**, 3210-3212.
- Simpson JT (2014) Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, **30(9)**, 1228-1235.
- Smeds L, Künstner A (2011) ConDeTri-a content dependent read trimmer for Illumina data. *PLoS One*, **6(10)**, e26314.
- Spielman D, Brook BW, Frankham R (2004) Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Science USA*, **101(42)**, 15261-15264.
- Vertebrate Genomes Project (2018) *Strigops habroptilus*. Retrieved from [https://vgp.github.io/genomeark/Strigops\\_habroptilus](https://vgp.github.io/genomeark/Strigops_habroptilus)
- Sutton J, Helmkampf M, Steiner C, *et al.* (2018) A high-quality, long-read de novo genome assembly to aid conservation of Hawaii's last remaining crow species. *bioRxiv*, 349035.
- Szulkin M, Bierne N, David P (2010) Heterozygosity-fitness correlations: a time for reappraisal. *Evolution*, **64(5)**, 1202-1217.
- Tamazian G, Dobrynin P, Krashenninnikova K, Komissarov A, Koepfli KP, O'Brien SJ (2016) Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. *GigaScience*, **5(1)**, 38.
- Taylor HR (2015) The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and Evolution*, **5(15)**, 3140-3150.
- Taylor HR, Dussex N, van Heezik YJ (2017) Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation*, **10**, 231-242.
- Tigano A, Sackton TB, Friesen VL (2018) Assembly and RNA-free annotation of highly heterozygous genomes: The case of the thick-billed murre (*Uria lomvia*). *Molecular Ecology Resources*, **18(1)**, 79-90.
- Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. *Nature Biotechnology*, **27(5)**, 455.
- Waldron A, Mooers AO, Miller DC, *et al.* (2013) Targeting global conservation funding to limit immediate biodiversity declines. *Proceedings of the National Academy of Science USA*, **110(29)**, 12144-12148.



- Wallis G (1999) *Genetic status of New Zealand black stilt (*Himantopus novaezelandiae*) and impact of hybridisation*. Wellington, New Zealand: Department of Conservation.
- Warren WC, Clayton DF, Ellegren H, *et al.* (2010) The genome of a songbird. *Nature*, **464(7289)**, 757.
- Waterhouse RM, Seppey M, Simão FA, *et al.* (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, **35(3)**, 543-548.
- Westbury MV, Hartmann S, Barlow A, *et al.* (2018) Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. *Molecular Biology and Evolution*, **35(5)**, 1225-1237.
- Willoughby JR, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, DeWoody JA (2015) The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology*, **24(1)**, 98-110.
- Zdobnov EM, Tegenfeldt F, Kuznetsov D, *et al.* (2016) OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, **45(D1)**, 744-749.
- Zerbino D, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18(5)**, 821-829.
- Zhang G (2015) Genomics: bird sequencing project takes off. *Nature*, **522(7554)**, 34.
- Zhang G, Li B, Li C, Gilbert MTP, Jarvis ED, Wang J (2014) Comparative genomic data of the Avian Phylogenomics Project. *GigaScience*, **3(1)**, 26.
- Zhang G, Li C, Li Q, *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346(6215)**, 1311-1320.

## Chapter 5: Comparing available tools for estimating relatedness and making pairing recommendations.

### 5.1: Preamble

This thesis has provided tools for conservation researchers to effectively and efficiently transition towards genomic technologies (**Chapters 3 & 4**), which offer exciting opportunities for informing management of critically-endangered species. As discussed in **Chapter 1**, one approach for the management of diversity within threatened populations is conservation breeding, where individuals are strategically paired and bred to minimise inbreeding and mean kinship (Ballou *et al.* 2010). While pedigrees are often the tool of choice for making these pairing recommendations, there are caveats to this approach which hinder its accuracy, including the assumption that all founders are unrelated, as highlighted in **Chapter 2**. When populations are poorly pedigreed (i.e., are  $\geq$  five generations deep, contain missing information, or individuals of unknown parentage; Balloux *et al.* 2004; Rudnick & Lacy 2008), an alternative approach for making pairing decisions is to use empirical estimates of relatedness via microsatellites or genome-wide SNPs to avoid inbreeding (Pemberton 2008; Speed & Balding 2015). To our knowledge, no one has compared pedigree-, microsatellite-, and SNP-based estimates of relatedness and downstream pairing recommendations for conservation breeding programmes. Therefore, the aim of this chapter, currently formatted as a manuscript In Review at *Evolutionary Applications*, is to compare these approaches using critically-endangered kakī and kākāriki karaka as Proof-of-Concept. We anticipate these results will underscore the differences

between these approaches and which will be effective and efficient for making pairing recommendations in conservation breeding programmes worldwide.

The Supplemental Materials for this chapter found in **Appendix G** and all scripts for SNP discovery, filtering, and analysis found in **Appendix H**. This chapter is also available as a pre-print on BioRxiv: <https://www.biorxiv.org/content/10.1101/721118v1>.

### *5.1.1: Contribution statement*

This research was led by me, with input from my supervisors Tammy Steeves and Anna Santure, with the research concept developed in collaboration with all co-authors. I completed all DNA extractions and microsatellite work presented here. I led the development of the kakī pedigree, and entered then edited it in PopLink with assistance from Liz Brown, Simone Cleland, and Cody Thyne (**Chapter 2**). The kākāriki karaka pedigree was developed by Anne Richardson, and I entered and edited it in PopLink with assistance from Anne Richardson (**Chapter 2**). The kākāriki karaka genome was assembled by Roger Moraga. I led SNP discovery for kakī and kākāriki karaka, with guidance from Roger Moraga. I conducted all statistical analyses, prepared all figures, and I led all drafts of this manuscript, incorporating edits from all coauthors.



**Stephanie J. Galla**

## 5.2: Comparison of pedigree, genetic, and genomic estimates of relatedness for informing pairing decisions in two critically-endangered birds: Implications for conservation breeding programmes worldwide.

Authors: Stephanie Galla<sup>1\*</sup>, Roger Moraga<sup>2</sup>, Liz Brown<sup>3</sup>, Simone Cleland<sup>3</sup>, Marc P. Hoepfner<sup>4</sup>, Richard F. Maloney<sup>5</sup>, Anne Richardson<sup>6</sup>, Lyndon Slater<sup>7</sup>, Anna W. Santure<sup>8</sup>, Tammy E. Steeves<sup>1</sup>

<sup>1</sup> School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

<sup>2</sup> Tea Break Bioinformatics, Ltd., Palmerston North, New Zealand

<sup>3</sup> New Zealand Department of Conservation, Twizel, New Zealand

<sup>4</sup> Institute for Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany

<sup>5</sup> New Zealand Department of Conservation, Christchurch, New Zealand

<sup>6</sup> The Isaac Conservation and Wildlife Trust, Christchurch, New Zealand

<sup>7</sup> New Zealand Department of Conservation, Rangiora, New Zealand

<sup>8</sup> School of Biological Sciences, University of Auckland, Auckland, New Zealand

\* Corresponding author, sgalla32@gmail.com

**Abstract:** Conservation management strategies for many highly threatened species include conservation breeding to prevent extinction and enhance recovery. Pairing decisions for these conservation breeding programmes can be informed by pedigree data to minimise relatedness between individuals in an effort to avoid inbreeding, maximise diversity, and maintain evolutionary potential. However, conservation breeding programmes struggle to use this approach when pedigrees are shallow or incomplete. While genetic data (i.e., microsatellites) can be used to estimate relatedness to inform pairing decisions, emerging evidence indicates this approach may lack precision in genetically depauperate species, and more effective estimates will likely be obtained from genomic data (i.e., thousands of genome-wide single nucleotide polymorphisms, or SNPs). Here, we compare relatedness estimates using pedigrees, microsatellites, and SNPs from whole genome resequencing approaches and subsequent pairing decisions in two critically endangered birds endemic to New Zealand: kakī/black stilt (*Himantopus novaezelandiae*) and kākāriki karaka/orange-fronted parakeet (*Cyanoramphus malherbi*). Our findings indicate that SNPs provide more

precise estimates of relatedness than microsatellites when assessing empirical parent-offspring and full sibling relationships. Further, our results show that relatedness estimates and subsequent pairing recommendations using *PMx* are most similar between pedigree- and SNP-based approaches. These combined results indicate that in lieu of robust pedigrees, SNPs are an effective tool for informing pairing decisions, which has important implications for many poorly pedigreed conservation breeding programmes worldwide.

*Keywords: Conservation genetics, conservation genomics, relatedness, conservation breeding, pairing recommendations, PMx.*

### 5.2.1: Introduction

In order to recover the world's rarest species, a multifaceted approach is needed to address the factors that cause species decline and those that promote species resilience (Soulé 1985; Jamieson 2015; Grueber *et al.* 2019). A critical facet of threatened species recovery is genetic management (Frankham 2005; O'Grady *et al.* 2006; Spielman *et al.* 2004), including conservation breeding, where breeding individuals in intensively managed captive populations are paired to minimise inbreeding and maximise genetic diversity in an effort to maintain evolutionary potential (Ballou & Lacy 1995; Ballou *et al.* 2010; Giglio *et al.* 2012; de Villemereuil *et al.* 2019). In these conservation breeding programmes, offspring may remain in captivity as an insurance population (e.g., the Tasmanian devil, *Sarcophilus harissii*, Hogg *et al.* 2015) or they may be translocated to the wild (e.g., California condor, *Gymnogyps californianus*; Walters *et al.* 2010) to enhance recovery efforts. In addition to demographic considerations (e.g., Slotta-Bachmayr *et al.* 2004; Tenhumberg *et al.* 2004; Moore *et al.* 2012), current best practice for making pairing decisions in conservation breeding programmes is to use available ancestry data from multigenerational pedigrees to

estimate kinship - a metric of pairwise coancestry or relatedness - between all living individuals in a population (Lacy 1995; Ballou & Lacy 1995; Ballou *et al.* 2010; Ivy & Lacy 2012). Individuals are paired to minimise mean kinship (i.e., average pairwise relatedness among all others in the population, including oneself), which has been shown to maximise founder representation and minimise inbreeding over time (Ballou & Lacy 1995; Lacy 2012; Willoughby *et al.* 2015).

Pedigrees are the tool of choice for estimating relatedness in conservation breeding programmes, including hundreds managed by the worldwide zoo and aquarium community (e.g., the Association of Zoos and Aquariums, or AZA; Long *et al.* 2011; Hammerly *et al.* 2016; Jiménez-Mena *et al.* 2016). Still, there are inherent assumptions that, when violated, hinder pedigree accuracy. For example, pedigrees assume no variance in founder relationships (i.e., all founders are equally unrelated; Ballou 1983), which is unlikely for many highly threatened wild populations, given most have experienced one or more historical population bottlenecks and founders sourced from these remnant wild populations will have variance in relatedness values (i.e., some founders will be more closely related than others; Bergner *et al.* 2014; Hogg *et al.* 2018). Simulation studies suggest that complete pedigrees with substantial depth (> 5 generations recorded) are robust enough to reflect true relatedness and inbreeding estimates despite violating this assumption, when a mean kinship approach is used for pairing (Balloux *et al.* 2004; Pemberton 2004; Rudnick & Lacy 2008). However, in many conservation breeding programmes, wild founders are routinely sourced to supplement captive populations (e.g., kākāriki karaka, *Cyanoramphus malherbi*, this manuscript) and to reduce the risk associated with adaptation to captivity (Frankham 2008). Under these circumstances, the assumption that there is no variance in founder relationships can be repeatedly violated, leading to

significant underestimation of relatedness and inbreeding (Hogg *et al.* 2018). In addition to these caveats, many intensively managed populations are poorly pedigreed, meaning these pedigrees contain missing information (i.e., unknown parentage due to matings that include unidentified individuals or extra-pair parentage; Lacy 2009; Bérénos *et al.* 2014; Pemberton 2008; Putnam & Ivy 2014) or record keeping errors (e.g., Hammerly *et al.* 2016).

Even when pedigrees are of high depth, have no missing information, and contain no errors, expected relatedness between individuals can differ from realised relatedness, as pedigrees are based on probabilities as opposed to empirical estimates of genome sharing (Hill & Weir 2012; Kardos *et al.* 2015; Speed & Balding 2015; Willoughby *et al.* 2015). Based on Mendelian inheritance, pedigrees estimate the probability that two alleles, one chosen at random from each of two individuals, are identical by descent (IBD) from a parent or common ancestor (Ballou 1983; Lacy 1995). When using a pedigree, the relatedness coefficient ( $R$ ) for parents and offspring, as well as for full siblings, is 0.5 when inbreeding is not present, indicating each pair shares 50% of their genomic information. While parents do contribute roughly 50% of their genomic information to their gametes, the combined effects of recombination, independent assortment, and random fertilisation can lead to a larger range of realised relatedness between full siblings (Hill & Weir 2011, 2012; Speed & Balding 2015). For example, a simulation study in humans revealed that realised relatedness between full siblings could range anywhere from 0.37-0.61 (Visscher *et al.* 2006), however this range can vary depending on the genome architecture of the species in question (e.g., number and size of chromosomes and the frequency and location of recombination events; Hill & Weir 2011; Kardos *et al.* 2015; Ulrich *et al.* 2017).

An alternative approach for populations lacking robust pedigrees is to use genetic-based estimates of pairwise relatedness to inform pairing decisions (Slate *et al.* 2004;

Pemberton 2004, 2008; Attard *et al.* 2016; Premachandra *et al.* 2019). This approach typically uses 8-30 microsatellite markers and empirical allele frequencies to estimate the probability that shared alleles are IBD from a common ancestor (Speed & Balding, 2015). To date, numerous conservation breeding programmes have used a genetic approach to inform pairing recommendations, repair studbooks, and resolve unknown parentage assignments, including programmes for the near-threatened Parma wallaby (*Macropus parma*; Ivy *et al.* 2009), the vulnerable Jamaican yellow boa (*Epicrates subflavus*; Tzika *et al.* 2009), the critically endangered Anegada iguana (*Cyclura pinguis*; Mitchell *et al.* 2011), and the critically endangered Attwater's prairie-chicken (*Tympanuchus cupido attwateri*; Hammerly *et al.* 2016; Hammerly *et al.* 2013). While some empirical research indicates that a large and diverse panel of microsatellites produces diversity estimates that are representative of genome-wide diversity and can be more useful than shallow pedigrees (e.g., Forstmeier *et al.* 2012), more recent simulation studies indicate that microsatellites provide less precision for relatedness and inbreeding, particularly in genetically depauperate endangered species where allelic diversity is low (i.e., < 4 alleles per locus in the founding population; Robinson *et al.* 2013; Taylor 2015; Taylor *et al.* 2015). While the use of larger panels of diverse microsatellites may circumvent this issue for some species (e.g., Bergner *et al.* 2014; Gooley *et al.* 2017), one simulation study for genetically depauperate endangered species shows that little precision is gained beyond 40 microsatellites, leading to inaccurate estimates of relatedness (Taylor *et al.* 2015). Recent studies argue that a better indication of genome-wide diversity can be obtained from genomic-based estimates of relatedness based on large numbers of genome-wide single nucleotide polymorphisms (i.e., SNPs; Knief *et al.* 2015; Taylor 2015; Taylor *et al.* 2015).



Given the decreasing cost of high-throughput sequencing (Hayden 2014) and the increasing amount of genomic resources readily available for non-model species (Galla *et al.* 2019), producing thousands of SNPs is now possible for many highly threatened species and provides an exciting opportunity for use in conservation breeding programmes (Galla *et al.* 2016; He *et al.* 2016). Indeed, there are several recent examples of genome-wide SNPs being used for relatedness in conservation, ecology, and evolution (e.g., De Fraga *et al.* 2017; Escoda *et al.* 2017), with some studies indicating that genome-wide SNPs provide greater precision in estimating relatedness and inbreeding compared to robust pedigrees (Santure *et al.* 2010; Kardos *et al.* 2015; Wang 2016) or microsatellites (Keller *et al.* 2011; Li *et al.* 2011; Béréños *et al.* 2014; Hellmann *et al.* 2016; Attard *et al.* 2018; Thrasher *et al.* 2018; Lemopoulos *et al.* 2019).

To our knowledge, no study has compared pedigree-, genetic-, and genomic-based approaches for estimating relatedness to inform pairing decisions for conservation breeding programmes, despite there being over 350 vertebrates worldwide that are captive bred for release to the wild (Smith *et al.* 2011). Here, we evaluate these three approaches using two critically endangered birds endemic to Aotearoa New Zealand — the kakī/black stilt (*Himantopus novaezelandiae*) and kākāriki karaka/orange-fronted parakeet (*Cyanoramphus malherbi*) — as Proof-of-Concept. Kakī and kākāriki karaka are excellent candidates for this research as both have active conservation breeding programmes, as well as multigenerational pedigrees (this study), microsatellite panels (Steeves *et al.* 2008; Andrews *et al.* 2013) and genomic resources including species-specific reference genomes and whole genome resequencing data (Galla *et al.* 2019; this study). In addition, because captive breeding pairs for both species are housed in separate enclosures and all offspring are

intensively managed, kakī and kākāriki karaka present an excellent opportunity to examine relatedness in known family groups.

Once found on both main islands of Aotearoa, kakī experienced significant population declines throughout the 20th century due to introduced mammalian predators (e.g., feral cats, *Felis catus*; stoats, *Mustela erminea*; and hedgehogs, *Erinaceus europaeus*) along with braided river habitat loss and degradation (Sanders & Maloney 2002). Today, an estimated 129 kakī are largely restricted to braided rivers of Te Manahuna/The Mackenzie Basin (Department of Conservation, *personal comm.*; Figure 5.1A) and recovery efforts include a conservation breeding programme that was initiated in the early 1980's (Reed 1998). In addition to breeding birds in captivity, the kakī recovery programme also collects eggs from intensively monitored wild nests and rears them in captivity before wild release. Similar to kakī, kākāriki karaka were also once found on both main islands of New Zealand and experienced population declines in the 19th and 20th centuries due to introduced mammalian predators (e.g., brushtail possums, *Trichosurus vulpecula*; rats, *Rattus rattus* and *R. norvegicus*; and stoats) and habitat loss (Kearvell & Legault 2017). Today, breeding populations of an estimated 100-300 kākāriki karaka are restricted to beech (*Nothofagus spp.*) forests in three North Canterbury Valleys (the Hawdon, Hurunui, and Poulter) and to Oruawairua/Blumine Island in the Marlborough Sounds (Department of Conservation, *unpublished data*; Figure 5.1B). Recovery efforts include a conservation breeding programme initiated in 2003, with founders sourced from the Poulter, Hawdon, and Hurunui Valleys. In most instances, offspring from pairings are released to the Hurunui Valley for wild supplementation. More recently, offspring are also released into the Poulter Valley to encourage pairing with the remaining birds from an extremely small remnant wild population (Department of Conservation, *personal comm.*). Eggs from these pairs are

harvested, brought into captivity, and fostered under surrogate birds, with hatchlings incorporated into the conservation breeding programme. Kakī belong to the Order Charadriiformes and are relatively long-lived braided river specialists that breed predictably within the bounds of a spring and summer season (Pierce 2013). In contrast, kākāriki karaka belong to the Order Psittaciformes and are relatively short-lived beech forest specialists capable breeding year round, with prolific breeding periods associated with food abundance (e.g., beech forest masting events; Kearvell & Legault 2017).

Here, we compare relatedness estimates from pedigree, microsatellites, and genome-wide SNPs using known parent-offspring and full sibling relationships. We then compare pairing recommendations among these three approaches to assess how each translates to effective conservation management. Given that kakī and kākāriki karaka represent two taxonomically distinct bird species with different life history strategies, we anticipate the results of our research may be applicable to the wider conservation breeding community.

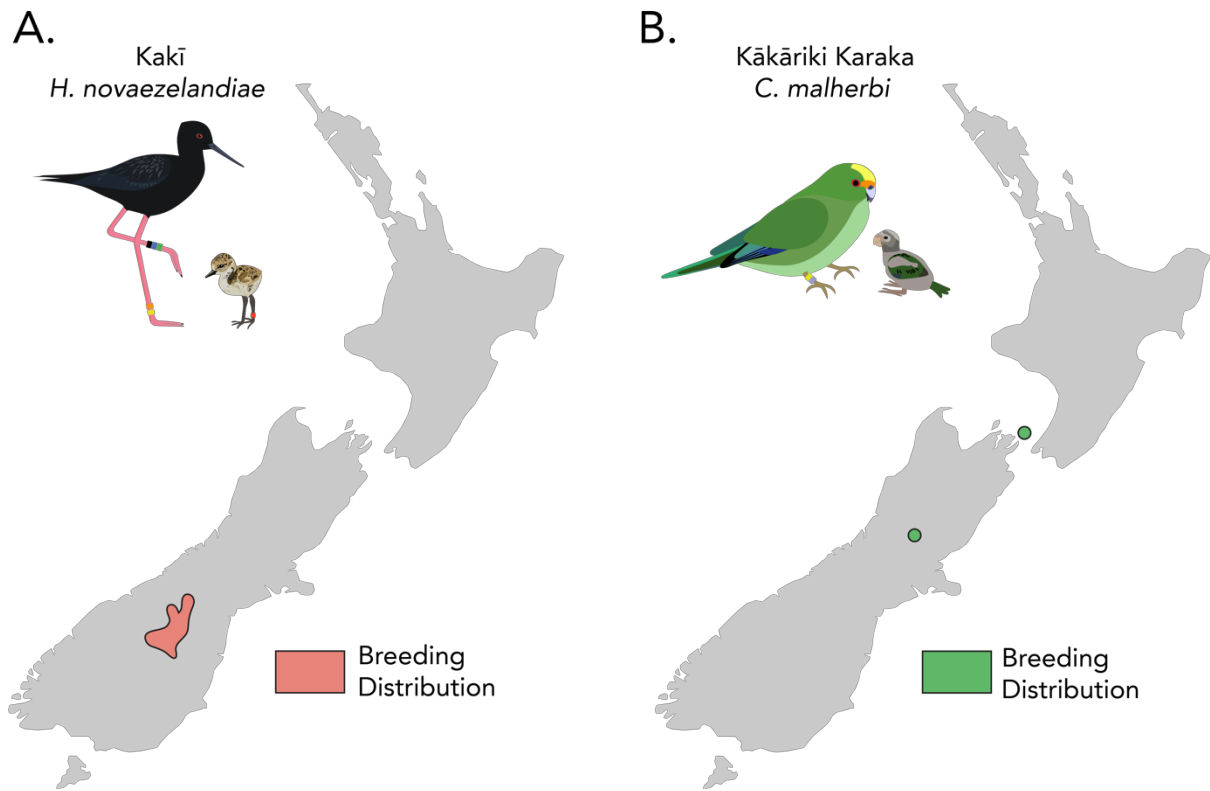


Figure 5. 1: Current breeding distributions of wild kakī (A) and kākāriki karaka (B) in Aotearoa.

## 5.2.2: Materials and methods

### 5.2.2.1: Sample Collection and DNA Extraction

Animal ethics approval for this project has been granted by the New Zealand Department of Conservation (i.e., DOC; permit number AEC 283). Captive kakī and kākāriki karaka are managed by DOC at two facilities in Aotearoa: the DOC Kakī Management Centre in Twizel and Isaac Conservation and Wildlife Trust in Christchurch. Kakī used in this study are 36 individuals sampled between 2014-2017, including 24 individuals from six captive family groups and 12 individuals from wild parents that represent diverse lineages based on the pedigree. Kākāriki karaka sampled in this study are 36 individuals sampled between 2015-2019, including individuals from eight captive family groups and one wild individual from the Poulter Valley of North Canterbury (Table 5.1).

Blood, feather, or tissue samples were sampled from each bird during routine health checks by DOC and Isaac Conservation and Wildlife Trust staff and immediately transferred into 95% molecular grade ethanol and stored at -80°C. High quantity and quality DNA was extracted using a lithium chloride chloroform extraction method (Galla et al. 2019) at the University of Canterbury School of Biological Sciences. Extractions were assessed for quality by running 2µL of DNA on a 2% agarose gel. A Qubit® 2.0 Fluorometer (Fisher Scientific) was used for DNA quantification.

Familial relationships are known for all samples collected, as they were sampled from birds of known provenance in captive conditions. However, to verify that no sample was mislabelled during genetic and genomic processing, parent-offspring relationships were verified through an allele mismatch exclusion analysis (Jones & Ardren 2003) using microsatellite panels previously developed for kakī (Steeves *et al.* 2008) and kākārīki karaka (Andrews 2013), with a maximum allowed mismatch of one allele at one locus (see *Microsatellite data* below). Family groups were further verified by clustering genome-wide SNP relatedness values calculated using the KGD method (Dodds *et al.* 2015) using principal component analysis and visualisation of family groups using the TensorFlow Embedding Projector (Smilkov *et al.* 2016; *data not shown*).

Table 5. 1: Family group sampling strategy used in this study.

Species	No. Sampled	No. Family Groups	No. Parent-Offspring Relationships	No. Sibling Relationships
Kakī	36	6	24	7
Kākārīki Karaka	36	8	52	48

### 5.2.2.2: Pedigree-based Relatedness

Multigenerational pedigrees were constructed for kakī and kākāriki karaka by entering studbook information (i.e., hatch date, sex, parentage, and status) into the programme PopLink v. 2.5.1 (Faust *et al.* 2018). Sex for all individuals was verified using molecular markers 2550F/2718R (Fridolfsson & Ellegren 1999) for kakī and P2/P8 (Griffiths *et al.* 1998) for kākāriki karaka, with PCR products run on a 3% agarose gel for visual characterisation, with positive controls included. Due to the short distance between P2/P8 alleles on the Z and W chromosomes in kākāriki karaka (Robertson & Gemmell 2006), 2µL of PCR products using a tagged forward primer were combined with 11.7µL formamide and 0.3µL of Genescan™ LIZ® 500 size standard (Applied Biosystems) and genotyped on an ABI 3739xl (Applied Biosystems), with alleles manually scored using GeneMarker v. 2.2 (State College, PA, USA). Inconsistencies in pedigrees were identified using the validation tool in PopLink and corrected using observations by DOC and the Isaac Conservation and Wildlife Trust. Pairwise estimates of kinship and inbreeding were produced using the programme PMx v. 1.6.20190628 (Lacy *et al.* 2012), selecting only the individuals used in this study ( $n = 36$  in kakī and kākāriki karaka) and treating all unknown individuals in the pedigree as wild founders. In order to produce direct comparisons of pairwise relatedness coefficients ( $R$ ) between pedigree, genetic and genomic data,  $R$  was calculated from pedigree kinship data using  $R(xy) = 2 * f(xy) / \sqrt{(1+F_x)(1+F_y)}$ . In this formula,  $f(xy)$  is the kinship between two individuals ( $x$  and  $y$ ) and  $F_x$  and  $F_y$  are the inbreeding coefficients of individuals  $x$  and  $y$  (Crow & Kimura 1970).

### 5.2.2.3: Genetic and Genomic-based Relatedness

#### 5.2.2.3.1: Microsatellite data

Microsatellite loci ( $n = 8$ ) previously described for kakī were amplified using PCR protocols by Steeves *et al.* (2008). Microsatellite loci ( $n = 17$ ) designed for kākāriki karaka and one locus (*Cfor0809*) for Forbe’s parakeet (*C. forbesi*) that cross-amplified in kākāriki karaka were amplified using PCR protocols by Andrews *et al.* (2013) and Chan *et al.* (2005), respectively. Samples were prepared for genotyping by adding 0.5 µl of PCR product to 11.7µl formamide and 0.3µl Genescan™ LIZ® 500 size standard (Applied Biosystems) and were genotyped on either a 3130xl or 3730xl Genetic Analyser (Applied Biosystems). Chromatograms were visualised using GeneMarker v. 2.2 (State College, PA, USA). To avoid bias by potential dye shifts (Sutton, Robertson, & Jamieson, 2011), peaks were scored manually. The number of alleles and standard estimates of per locus diversity — including expected heterozygosity ( $H_o$ ) and observed heterozygosity ( $H_E$ ) — were produced using GenAlEx v. 6.5 (Peakall & Smouse 2006; Smouse & Peakall 2012). Tests for deviations from Hardy-Weinberg and linkage disequilibrium for these loci using samples that are representative of larger kakī and kākāriki karaka populations can be found in Steeves *et al.* (2008; 2010) and Andrews (2013) respectively. For kākāriki karaka, only eight of the 18 microsatellite markers previously described were polymorphic in this study and these eight loci were used in all downstream analyses.

Genetic-based  $R$  estimates were produced in the programme COANCESTRY v. 1.0.1.9 (Wang 2011). COANCESTRY offers seven different estimators of relatedness, and to choose the most appropriate estimator for the kakī and kākāriki karaka microsatellite data sets, we employed the simulation module within COANCESTRY using allele frequencies, missing data, and error rates from our microsatellite data sets. To produce dyads that represent the relationships and degree of inbreeding found within kakī and kākāriki karaka, we used R package ‘identity’ (Li 2010) to generate 10,879 dyads for kakī and 1,484 dyads for kākāriki

karaka based on the pedigrees of both species. The frequency of each unique dyad in the kakī and kākāriki karaka data sets were scaled to create 1,000 dyads for each set that are representative of relationships between individuals used in this study. The COANCESTRY simulations were conducted using allele frequencies, error rates, and missing data rates from each microsatellite data set, with settings changed to account for inbreeding. The triadic likelihood approach (Wang 2007) was selected given it had the highest Pearson's correlation with 'true' relatedness for both data sets (see Appendix G for details). This approach is also preferred, as it is one of the few estimators that accounts for instances of inbreeding (Wang 2007).

To estimate  $R$  with our genetic data set, COANCESTRY programme parameters were set to account for inbreeding, with the number of reference individuals and bootstrapping samples set to 100.

#### 5.2.2.3.2: Genomic Data

##### 5.2.2.3.2.1: Reference Genomes

A reference genome for kakī has already been assembled (Galla *et al.* 2019) and was used in this study. To assemble a *de novo* reference genome for kākāriki karaka, a paired-end library was prepared at the Institute of Clinical Molecular Biology (IKMB) at Kiel University using the Nextera™ DNA Flex Library Prep Kit according to manufacturer specifications and sequenced on an Illumina NovaSeq™ 6000 with 2 x 150 bp reads at a depth of approximately 70x.

FastQC v. 0.11.8 (Andrews 2010) was used to evaluate the quality of the raw Illumina data and assess potential sample contamination. Initial read trimming was performed using TrimGalore v. 0.6.2 (Krueger 2019) and Cutadapt v. 2.1 (Martin 2011) with an end trim quality of 30, a minimum length of 54, and using the --nextseq two-color chemistry option.



A median Phred score of 20 was also used for initial read trimming to remove obvious data errors; however, it should be noted that the assembly programmes used here (i.e., Meraculous-2D v. 2.2.10 and MaSuRCA v. 3.2.9; see below) have their own error corrections embedded in their respective pipelines. Kmer analyses were performed using Jellyfish v. 2.2.10 (Marçais & Kingsford, 2011) prior to assembly to assess heterozygosity and contamination. Two genome assembly programmes were tested for assembly performance: Meraculous-2D v. 2.2.5.1 (Chapman *et al.* 2011) and MaSuRCA v. 3.2.9 (Zimin *et al.* 2013). Meraculous was run using trimmed reads in diploid mode 1, with all other assembly parameters set to default. MaSuRCA was run using untrimmed reads, as it incorporates its own error correction pipeline. MaSuRCA parameters adjustments include a grid batch size of 300,000,000, the longest read coverage of 30, a Jellyfish hash size of 14,000,000,000, and the inclusion of scaffold gap closing; all other parameters were set to default for non-bacterial Illumina assemblies. The final assembly using the Meraculous pipeline was more fragmented (i.e., an N50 of 28.5 kb with 67,046 scaffolds > 1 kb), while the MaSuRCA genome was less fragmented (i.e., an N50 of 107.4 kb with 66,212 scaffolds > 1 kb) but contained possible artefacts due to heterozygosity (i.e., tandem repeats flanking short stretches of “N”s). To correct for these issues, the Meraculous assembly was first aligned to the MaSuRCA assembly using Last v. 959 (Kielbasa *et al.* 2011), then alignments were filtered to find matches where the Meraculous assembly spans the entirety of the tandem repeat in the MaSuRCA scaffolds, but lacking the tandem repeat or stretch of “N”s (i.e., gaps). In those cases, the aligned sequence in the MaSuRCA scaffold was replaced with the Meraculous match. All compute requirements needed to assemble the kākārīki karaka genome are available in Appendix G.

#### 5.2.2.3.2.2: Whole-genome Resequencing

Kakī resequencing libraries were prepared at IKMB using a TruSeq® Nano DNA Library Prep kit following the manufacturer's protocol and were sequenced across 34 lanes of an Illumina HiSeq 4000. 24 individuals were sequenced at high coverage depth (approximately 50x) for an aligned study, and all others were sequenced at a lower coverage depth (approximately 10x). Kākāriki karaka libraries were prepared at IKMB using the Nextera™DNA Flex Library Prep Kit according to manufacturer specifications and sequenced across one lane of an Illumina NovaSeq™6000 at IKMB at a coverage depth of approximately 10x, with one individual sequenced at a depth of approximately 70x, which was additionally used for the reference genome (see above).

FastQC v. 0.11.4 and 0.11.8 (Andrews 2010) were used to evaluate the quality of the raw Illumina data for kakī and kākāriki karaka, respectively. Kakī resequencing reads were subsequently trimmed for the Illumina barcode, a minimum Phred quality score of 20, and a minimum length of 50bp using Trimmomatic v. 0.38 (Bolger *et al.* 2014). Because kākāriki karaka libraries were produced using different library preparation protocols and nextera chemistry, reads were trimmed using TrimGalore v. 0.6.2 (Krueger 2019) for nextera barcodes and two-colour chemistry, using a median Phred score of 20, end trim quality of 30, and a minimum length of 54. Prior to mapping, the kakī reference genome was concatenated to a single chromosome using the custom perl script 'concatenate\_genome.pl' (Moraga 2018a) for use in an aligned project that used both resequencing and genotyping-by-sequencing reads (see Galla *et al.* 2019). The kakī and kākāriki karaka reference genomes were indexed and resequencing reads were mapped using Bowtie2 v. 2.2.6 and v. 2.3.4.1 (Langmead & Salzberg 2012), respectively, with the setting *--very-sensitive*. Resulting SAM files were converted to BAM and were sorted using

Samtools v. 1.9. (Li *et al.* 2009). Read coverage and variant calling were performed using *mpileup* in BCFtools v. 1.9 (Li *et al.* 2009). The custom perl script 'split\_bamfile\_tasks.pl' (Moraga, 2018b) was used to reduce the computational time needed for *mpileup* by increasing parallelisation. SNPs were detected, filtered, and reported using BCFtools. Filtering settings were set to retain biallelic SNPs with a minor allele frequency (MAF) greater than 0.05, a quality score greater than 20, and a maximum of 10% missing data per site. After a series of filtering trials for each species (see Appendix G for details), depth for kakī was set to have an average mean depth greater than 10, while kākāriki karaka depth was set so that each site had a minimum depth of 5 and a maximum depth of 200. Resulting SNPs for both data sets were pruned for linkage disequilibrium using BCFtools with the  $r^2$  set to 0.6 and a window size of 1000 sites. Sites were not filtered for Hardy-Weinberg equilibrium, as the nature of these data sets (mostly family groups) violates the assumptions of random mating. Per site missingness, depth, and diversity — including proportion of observed and expected heterozygous SNP sites per individual, nucleotide diversity, and SNP density per kb — were evaluated in the final sets using VCFtools v. 1.9 (Danecek *et al.* 2011). Diversity statistics were calculated using polymorphic markers only.

#### 5.2.2.3.2.3: SNP-based Relatedness

To produce estimates of  $R$  using whole-genome SNPs, the programme KGD (Dodds *et al.* 2015) was used, as it was designed to estimate relatedness using reduced-representation and resequencing data while taking into account read depth. Pairwise  $R$  values derived from KGD were scaled so that self-relatedness for all individuals was equal to 1 using the formula  $MS = D \times MO \times D$  where  $MS$  is the scaled matrix,  $MO$  is the original matrix, and  $D$  is a diagonal matrix with elements  $D = 1/\sqrt{\text{diag}(MO)}$ . This scaling was performed to simplify downstream Mantel tests by creating a standardised diagonal value. This scaling was

maintained throughout all analyses, as the scaled approach better approximated parent-offspring relationships, while demonstrating minimal bias to downstream analyses (see Appendix G for details).

We evaluated the scaled KGD approach with other maker-based relatedness estimators, including the triadic likelihood approach (Wang 2007), the KING estimator (Waples *et al.* 2019), and the  $r_{xy}$  method (Hedrick & Lacy 2015), using parent-offspring relatedness as a benchmark for precision. We found that the scaled KGD approach estimates parent-offspring relatedness closer to 0.5 compared to other relatedness estimators, while still providing estimates that are significantly concordant with all other approaches in both kakī (Pearson's  $r = 0.80-0.96$ ,  $p < 0.001$ ) and kākāriki karaka (Pearson's  $r = 0.89-0.96$ ,  $p < 0.001$ ; see Appendix G for details).

#### **5.2.2.4: Comparison of Relatedness**

Mantel tests using the R-package *ape* (Paradis & Schliep 2018) were performed with 1000 iterations to determine whether pedigree, microsatellite, and SNP-based  $R$  were significantly correlated compared to a null distribution. Pearson's correlation coefficient ( $r$ ) was additionally calculated to provide an additional measure of concordance between approaches. While our relatedness data sets are non-parametric, Pearson's was used over non-parametric tests, such as rank correlations, as our pedigree and microsatellite data sets have an excess of tied values.

#### **5.2.2.5: Pairing Recommendations**

We used two complementary methods in PMx v. 1.6.20190628 (Lacy *et al.* 2012) to determine if pairing recommendations change using pedigree-, microsatellite-, and SNP-based approaches for estimating  $R$ . First, we used mate suitability index (MSI), which scores how valuable offspring of a potential pair would be by taking into account four factors:

deltaGD (i.e., the net positive or negative genetic diversity provided to the population), the difference of mean kinship values of the pair, the inbreeding coefficients of resulting offspring, and the extent of unknown ancestry (Ballou *et al.* 2001; Lacy *et al.* 2011). MSI scores scale from 1-6, with 1 being “very beneficial”, and 6 being “very detrimental”. An additional category denoted with a “-” indicates “very highly detrimental” pairings. Here, we assign this category with a numerical MSI score of 7. MSI scores provide a standardised approach for comparing pairing recommendations within and among species, including those based on the three approaches used in this study. However, Ballou *et al.* (2001) recommend caution when using automated pairing recommendations such as MSI in small and inbred populations. Thus, we also used mean kinship (MK) rank, which is known to perform well in small and inbred populations (Ballou & Lacy 1995; Rudnick & Lacy 2008). This approach ranks individuals from lowest to highest MK amongst males and females, thereby creating a list of individuals for pairing prioritisation.

For MSI score and MK rank analyses, only the individuals used in this study ( $n = 36$  for both kakī and kākārīki karaka) were selected for analysis. PMx settings were set to default, with the exception of treating all unknown individuals in the pedigree as wild (i.e., 100% analytics known in the pedigree) to minimise bias from unknown pedigree assignments. Pedigree-based MSI scores and MK ranks were produced using pedigree-based kinship, while pairing recommendations using microsatellites and SNPs were produced using coefficients of relatedness. These genetic and genomic estimates of relatedness were uploaded to PMx, which divides these values by two to create an empirical metric of kinship. These empirical values were weighted to 1 to produce MSI scores and MK ranks that relied only on empirical data.

Pearson's correlation ( $r$ ) was used to evaluate whether pairwise MSI scores and MK ranks between approaches were concordant. To test whether the distribution of MSI scores were statistically different from one another, we used a non-parametric Kruskal-Wallis test with Bonferroni correction and a Tukey Honest Significant Difference (HSD) test.

### 5.2.3: Results

#### 5.2.3.1: Pedigree-based Relatedness

This study has produced the first multigenerational pedigrees for two critically endangered endemic birds from Aotearoa. The kakī pedigree includes 2,481 wild and captive individuals recorded from 1977-present, with a pedigree depth ranging from 1-8 generations (3.35 average). The number of founders and founder genome equivalents in the kakī pedigree (94 and 12.9, respectively) is high relative to kākāriki karaka (16 and 12, respectively), with a higher degree of uncertain ancestry (i.e., there are no individuals of unknown parentage in the kākāriki karaka pedigree; Table 5.2). Pedigree-based  $R$  between the 36 focal kakī ranged from 0 to 0.56, with an average  $R$  of  $0.13 \pm \text{SD } 0.13$ . The average coefficient of relatedness between all known kakī parent-offspring was higher than the expected 0.5 contribution from each parent ( $0.52 \pm \text{SD } 0.02$ ), with averaged full sibling  $R$  of  $0.52 \pm \text{SD } 0.02$  (Figure 5.2). The kākāriki karaka pedigree includes 624 captive individuals from 2003-present, with a pedigree depth ranging from 1-5 generations (2.48 average). Pedigree-based  $R$  for the 36 focal kākāriki karaka ranged from 0 to 0.67, with an average  $R$  of  $0.19 \pm \text{SD } 0.18$ . Average  $R$  between all parent-offspring was  $0.52 \pm \text{SD } 0.03$ , with averaged full sibling  $R$  being  $0.51 \pm \text{SD } 0.02$  (Figure 5.2).

*Table 5. 2 Descriptive statistics based on pedigree data, as produced by PMx, including number of individuals, sex ratio (males:females), maximum age, gene diversity, number of founders, number of founder genome equivalents, average inbreeding, average mean kinship, average generation time, % ancestry and analytic known, and effective population size. \*The sex ratio for all pedigreed individuals for kakī is biased by a large number of individuals with unknown sex.*

Pedigree Statistic	All Pedigreed Individuals		Individuals in Study	
	Kakī	Kākāriki Karaka	Kakī	Kākāriki Karaka
No. Individuals	2481	618	36	36
Sex Ratio	0.27*	0.5	0.44	0.5
Max. Age (Years)	24	16.4	19.3	19.6
Gene Diversity	0.96	0.915	0.9112	0.886
No. Founders	94	16	29	12
Founder Genome Equivalents	12.9	12	5.6	4.4
Average Inbreeding	0.027	0.03	0.034	0.016
Average Mean Kinship	0.039	0.085	0.089	0.114
Average Generation Time	4.82	1.31	5.25	3.79
% Ancestry Known	55	100	58	100
% Analytic Known	100	100	100	100
Ne/N	0.103	0.072	0.353	0.541

### 5.2.3.2: Microsatellite Diversity and Relatedness

All eight microsatellite markers for kakī successfully amplified in all individuals used in this study. The number of alleles present across kakī loci ranged from 2-4 (average  $3.13 \pm$  SD 0.64; Table 5.3), with overall fewer alleles found here than reported in previous studies utilising these loci with more individuals (Steeves *et al.* 2010; Hagen *et al.* 2011). While eighteen microsatellite markers were amplified in kākāriki karaka, one was removed from this study for not successfully amplifying in more than 50% of individuals (locus *OFK56*) and nine were removed for being monomorphic (Table 5.3). The number of alleles among polymorphic loci ranged from 2-4 (average  $3.0 \pm$  SD 0.93), with overall fewer alleles found here than reported in previous studies (Andrews 2013; Andrews *et al.* 2013). Observed ( $H_O$ ) and expected ( $H_E$ ) heterozygosity for kakī (average  $H_O = 0.57 \pm$  SD 0.17, average  $H_E = 0.54 \pm$  SD 0.14) was higher than kākāriki karaka (average  $H_O = 0.43 \pm$  SD 0.23, average  $H_E = 0.43 \pm$  SD 0.20; Table 5.3).

Microsatellite-based  $R$  between all kakī used in this study ranged from 0 to 0.85, with an average  $R$  of  $0.16 \pm \text{SD } 0.19$ . Average  $R$  between all known kakī parent-offspring ( $0.44 \pm \text{SD } 0.13$ ) was below the minimum expected relatedness value of 0.5, with a larger standard deviation of  $R$  values compared to pedigree-based estimates. Averaged full sibling  $R$  ( $0.41 \pm \text{SD } 0.20$ ) also had a larger deviation around the mean compared to the microsatellite-based parent-offspring estimates (Figure 5.2).

Microsatellite-based  $R$  between all kākāriki karaka used in this study ranged from 0 to 0.84, with an average  $R$  of  $0.18 \pm \text{SD } 0.22$ . Similar to kakī, average  $R$  between all known kākāriki karaka parent-offspring relationships ( $0.47 \pm \text{SD } 0.19$ ) was below the minimum expected  $R$  value of 0.5, with a larger standard deviation of  $R$  values compared to pedigree-based estimates. Averaged full sibling  $R$  ( $0.49 \pm \text{SD } 0.21$ ) also had a larger deviation around the mean compared to microsatellite-based parent-offspring estimates (Figure 5.2).

*Table 5. 3: Descriptive statistics, including number of alleles, observed heterozygosity ( $H_O$ ), and expected heterozygosity ( $H_E$ ) for microsatellite loci used in this study. Loci from kākāriki karaka that were monomorphic (OFK12, OFK 19, OFK21, OFK26, OFK31, OFK33, OFK52, OFK56, OFK58, OFK61) are not included.*

Species	Locus	No. Alleles	$H_O$	$H_E$
Kakī	BS2	3	0.667	0.652
	BS9	3	0.611	0.59
	BS12	3	0.278	0.245
	BS13	2	0.528	0.5
	BS21	4	0.833	0.703
	BS27	4	0.667	0.551
	BS40	3	0.444	0.448
	BSdi7	3	0.556	0.596
Kākāriki Karaka	OFK9	4	0.472	0.477
	OFK41	4	0.75	0.702
	OFK50	3	0.444	0.513
	OFK54	4	0.722	0.574
	OFK55	2	0.278	0.346
	OFK60	2	0.222	0.239
	OFK62	2	0.083	0.08
	C for 809	3	0.44	0.52

### 5.2.3.3: Reference Genome Assembly, SNP Discovery, Diversity, and Relatedness Estimates



#### 5.2.3.3.1: Kākāriki Karaka Reference Genome Assembly

Reference genome library preparation and Illumina NovaSeq™ sequencing resulted in 584.47 million total reads for the kākāriki karaka genome. The final kākāriki karaka genome assembly was 1.15GB in length, which is within the range of most assembled avian genomes (e.g., Zhang *et al.* 2014). The final assembly had 66,212 scaffolds with a scaffold N50 of 107.4 kb. See Data Availability section for access information.

#### 5.2.3.3.2: SNP Discovery and Diversity

Library preparation and Illumina sequencing resulted in 6.07 billion total reads for kakī ( $168.69 \pm \text{SD } 65.32$  million reads). In addition to the individual used for the reference assembly, 3.64 billion total reads (average =  $103.92 \pm \text{SD } 29.76$  million reads) were produced for the additional 35 kākāriki karaka in this study. More SNPs were discovered during initial SNP discovery using kākāriki karaka than kakī, and more remained post filtering (Table 5.4). These filtered SNPs were used for all downstream analyses. Average missingness was low for both data sets (Table 5.4), but lower for kākāriki karaka than kakī, as kākāriki karaka had a hard minimum cut-off for depth during filtering that resulted in no missing data. Average depth for both data sets was relatively high (Table 5.4), with kakī having slightly higher average depth. Average diversity statistics (nucleotide diversity, and the average observed and expected SNP heterozygosity per individual post filtering) were similar in both species, with diversity in kakī being slightly higher. SNP density using the kakī data set was higher than the kākāriki karaka data set, indicating that discovered SNPs are closer in proximity in the kakī data set (Table 5.4).

Table 5. 4: Descriptive statistics, including number of SNPs pre- and post- filtering, average depth, average missingness, average nucleotide diversity ( $\pi$ )  $\pm$  SD, average proportion of observed heterozygous SNP sites ( $H_O$ )  $\pm$  SD, average proportion of expected heterozygous SNP sites ( $H_E$ )  $\pm$  SD, and average SNP density (number of SNPs per kilobase)  $\pm$  SD.

Species	No. of SNPs Pre-Filtering	No. of SNPs Post-Filtering	Average Depth	Average Missingness	Average $\pi$	Average $H_O$ of SNP Sites	Average $H_E$ of SNP Sites	SNP Density
Kakī	4,246,100	68,144	28.73 $\pm$ 10.29	0.002 $\pm$ 0.004	0.35 $\pm$ 0.14	0.40 $\pm$ 0.02	0.35 $\pm$ 0.00	0.58 $\pm$ 3.18
Kākāriki Karaka	22,435,128	90,949	25.1 $\pm$ 14.87	0.00 $\pm$ 0.00	0.33 $\pm$ 0.14	0.37 $\pm$ 0.02	0.33 $\pm$ 0.00	0.17 $\pm$ 0.42

#### 5.2.3.3.3: SNP-based Relatedness

SNP-based  $R$  between all kakī used in this study ranged from 0.13-0.61, with an average  $R$  of  $0.27 \pm \text{SD } 0.09$ . Similar to pedigree-based estimates, average  $R$  between all known kakī parent-offspring were slightly higher than the expected relatedness value of 0.5 with a small standard deviation relative to microsatellite-based estimates ( $0.54 \pm \text{SD } 0.03$ ). Averaged full sibling  $R$  also had a larger deviation around the mean ( $0.52 \pm \text{SD } 0.05$ ) than parent-offspring relationships (Figure 5.2).

SNP-based  $R$  between all kākāriki karaka used in this study ranged from 0.08-0.67, with an average  $R$  of  $0.30 \pm \text{SD } 0.12$ . Similar to pedigree-based estimates, average  $R$  between all known kākāriki karaka parent-offspring was slightly above the expected  $R$  value of 0.5 with a small standard deviation relative to genetic-based estimates ( $0.53 \pm \text{SD } 0.03$ ). Averaged full sibling relatedness also had a larger deviation around the mean ( $0.52 \pm \text{SD } 0.05$ ) compared to the pedigree-based estimates (Figure 5.2).

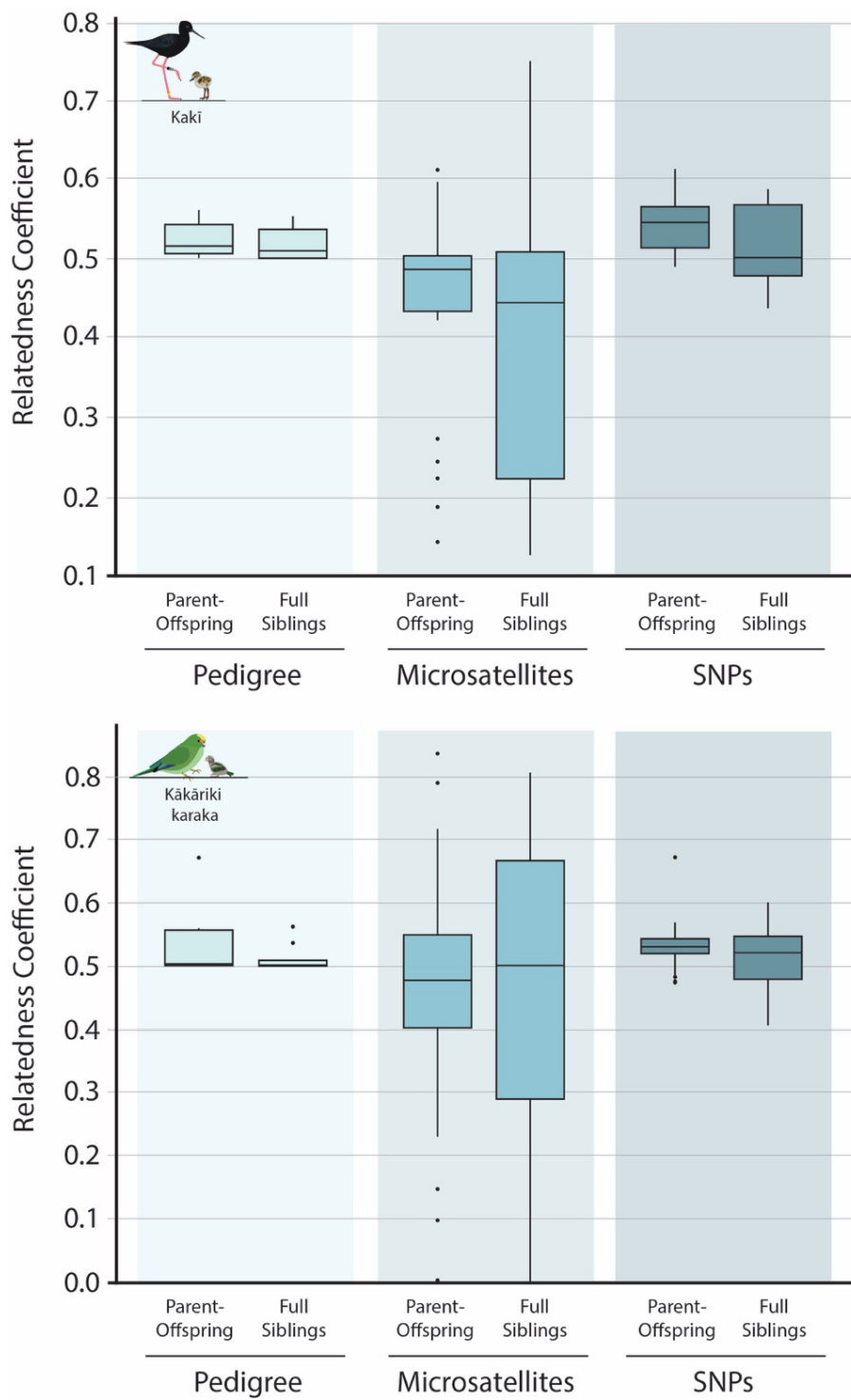


Figure 5. 2: Parent-offspring and full sibling relatedness values derived from pedigree- (pale blue), microsatellite- (medium blue), and SNP-based (dark blue) methods in kakī (top graph) and kākāriki karaka (bottom graph).

#### 5.2.3.4: Comparison of Relatedness Estimates and Pairing Recommendations

All kakī and kākāriki karaka  $R$  estimates using pedigree-, microsatellite-, and SNP-based approaches correlated with one another with high statistical significance (Mantel test,  $p < 0.001$ ; Pearson's Correlation,  $p < 0.001$ ; Figure 5.3). Of all the approaches, the correlation coefficient between pedigree- and SNP-based approaches was markedly higher than between other approaches, indicating that they are the most concordant (Figure 5.3).

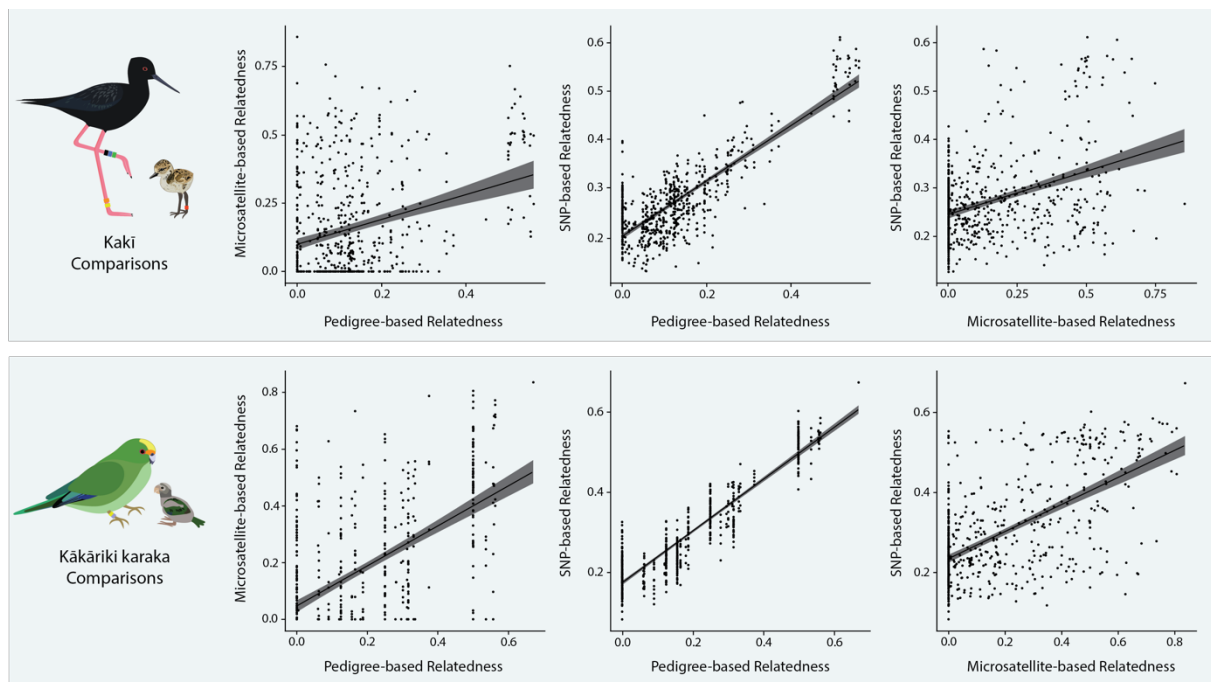


Figure 5. 3: Scatterplots showing relationships between pedigree-, microsatellite-, and SNP-based relatedness estimates in known family groups for kakī and kākāriki karaka. A trend line (black) and 95% confidence intervals (grey) are shown in each comparison..

Mate-suitability index (MSI) scores and MK ranks were calculated as an approximation for pairing recommendations derived from  $R$  estimates using the different approaches. Average pedigree-based MSI scores for kakī ( $4.46 \pm \text{SD } 1.59$ ) were lower on average than microsatellite-based scores ( $4.73 \pm \text{SD } 1.63$ ), but not significantly different from each other (Kruskal-Wallis test with Bonferroni correction,  $p = 0.2$ ). SNP-based MSI scores for kakī (average =  $5.67 \pm \text{SD } 1.39$ ) were significantly higher than pedigree- and microsatellite-based scores (Kruskal-Wallis test with Bonferroni correction,  $p < 0.001$ ), with

SNP-based scores providing the highest frequency of category 7 (i.e., very highly detrimental) pairings (Figure 5.4). While the distributions of MSI scores between each approach were different, each approach produced scores that correlated significantly with one another (Pearson's correlation,  $p < 0.01-0.001$ ). Similar to correlations between  $R$  estimates, correlation coefficients between pedigree and SNP-based MSI scores were highest. Of the 320 possible kakī pairings, 38% did not experience a change in MSI score value between pedigree- and SNP-based approaches; however, 20% of pairings experienced an MSI score change that was 3+ categories different. In 2% of pairings, pedigree-based MSI scores were categorised as a 1 (i.e., preferred pairing) while SNP-based MSI scores were categorised as a 7 (i.e., very highly detrimental). Correlations between MK ranks provided by the three approaches were significant between pedigree- and SNP- based approaches only (Pearson's  $r = 0.75$ ,  $p < 0.001$ ; see Figure G2 for details). Amongst pedigree- and SNP-based MK ranks, 64% of individuals experienced a minimal rank shift of 0-3 categories, 22% experienced a moderate rank shift of 4-7 categories, and 3% experienced a high rank shift of  $\geq 8$  categories.

Similar to kakī, average kākāriki karaka SNP-based MSI scores ( $5.64 \pm \text{SD } 1.47$ ) were significantly higher than pedigree ( $5.20 \pm \text{SD } 1.71$ ) and microsatellite ( $5.04 \pm \text{SD } 1.61$ ) scores (Kruskall-Wallis test with Bonferroni correction,  $p < 0.001$ ), while pedigree- and microsatellite-based scores did not significantly differ (Kruskall-Wallis test with Bonferroni correction,  $p = 0.67$ ). SNP-based scores provided the highest frequency of category 7 (i.e., very highly detrimental) pairings (Figure 5.4). Each approach also produced scores that correlated significantly with one another (Pearson's correlation,  $p < 0.001$ ), with the highest correlation coefficients seen between pedigree and SNP-based MSI scores (Pearson's  $r = 0.65$ ). Of the 324 possible pairings for kākāriki karaka, 59% did not experience a change in

MSI score value between pedigree- and SNP-based approaches; however, 9% of pairings experienced an MSI score change that was 3+ categories different. In 2% of pairings, pedigree-based MSI scores were categorised as a 1 (i.e., very beneficial) while SNP-based MSI scores were categorised as a 7 (i.e., very highly detrimental). Correlations between MK Rank were significant between pedigree- and SNP-based approaches (Pearson's  $r = 0.64$ ,  $p < 0.001$ ) and microsatellite- and SNP-based approaches (Pearson's  $r = 0.51$ ,  $p = 0.002$ ). Amongst pedigree- and SNP-based MK ranks, 53% of individuals experienced a minimal rank shift of 0-3 categories, 31% experienced a moderate rank shift of 4-7 categories, and 8% experienced a high rank shift of  $\leq 8$  categories.

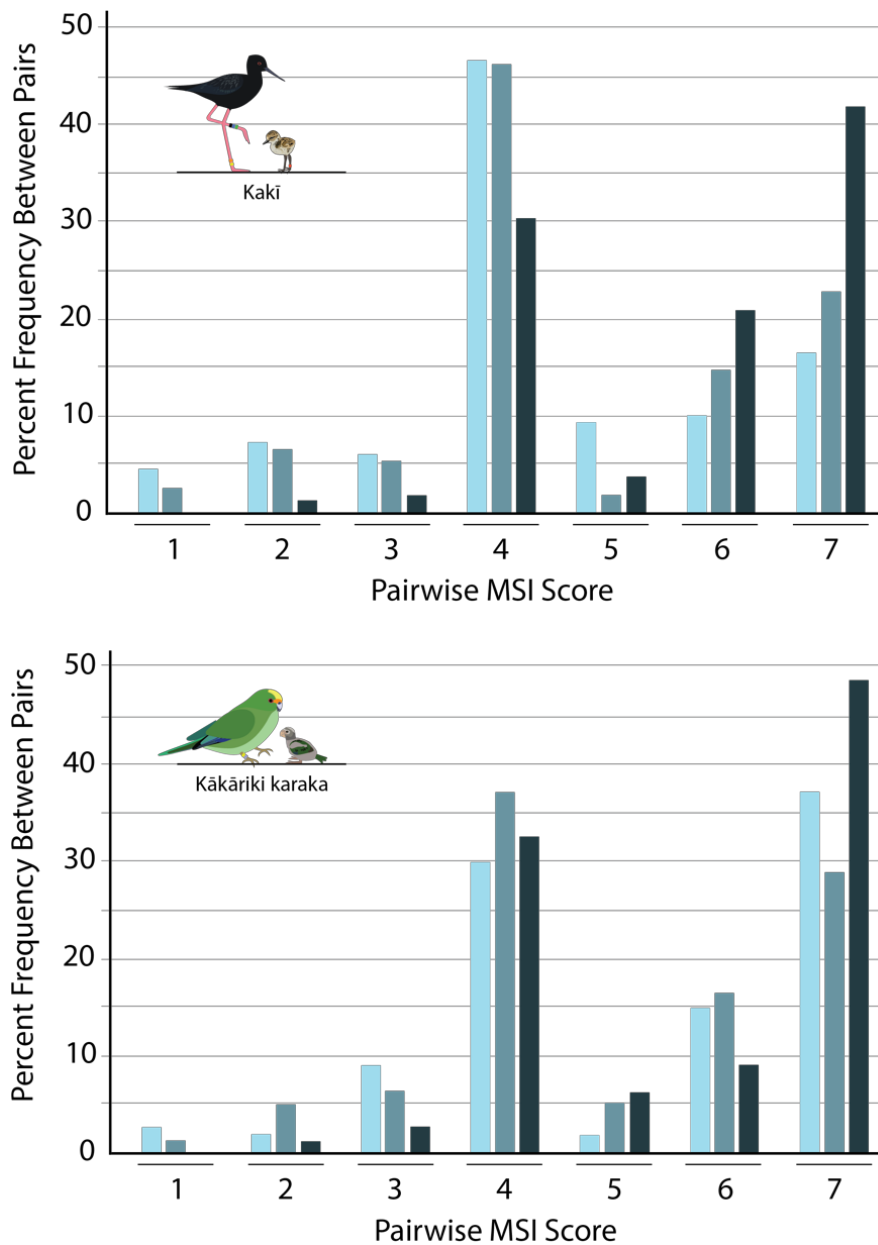


Figure 5. 4: Frequency of MSI scores using pedigree - (pale blue), microsatellite- (medium blue), and SNP-based (dark blue) kinship/relatedness values in kākī and kākāriki karaka.

#### 5.2.4: Discussion

This study is the first to compare pedigree-, microsatellite-, and SNP-based estimates of relatedness and subsequent pairing recommendations for conservation breeding programmes. The results indicate that microsatellites provide the least precision when estimating relatedness in known family groups, with pedigree- and SNP-based estimates providing higher precision and a much closer approximation of parent-offspring and full

sibling relatedness. Further, estimates of relatedness and downstream pairing recommendations using MSI scores and MK ranks are both more concordant when using pedigree- and SNP-based data sets compared to microsatellite-based data sets. Despite this, there were important differences in pairing recommendations between pedigree and SNP approaches, with SNP-based mate suitability index (MSI) scores being statistically higher than pedigree-based scores, and some substantial disagreements existing between the two sets of MSI scores and MK ranks. Together, this study provides insight into the differences between pedigree-, microsatellite-, and SNP-based approaches for making pairing recommendations and a pathway for estimating relatedness using genome-wide SNPs to inform pairing decisions in poorly pedigreed conservation breeding programmes worldwide.

#### **5.2.4.1: Relatedness Comparisons**

When producing empirical estimates of relatedness, simulations were performed to choose the most suitable estimator for microsatellites, and various programmes and filtering schemes were evaluated using SNPs to find an approach that best approximated known parent-offspring relationships. While different relatedness estimators and filtering schemes will result in different point estimates of relatedness, this study demonstrates an approach for producing relatedness estimates that are well suited for our particular data set.

Pedigree-based estimates of parent-offspring and full sibling relatedness approximated 0.5 for both kakī and kākāriki karaka (Figure 5.2), with some measures being slightly higher, which likely reflects intergenerational inbreeding. These results are consistent with expectations, as pedigrees are based on the probability of Mendelian inheritance, which postulates that first-order relationships (i.e., parents and offspring, and



siblings) share 50% of their genomic information (Lacy, 1995; Wright, 1922). We expect realised (i.e., empirical) parent-offspring relationships to also approximate 0.5, but a broader range of realised relatedness estimates among full siblings, as they may receive different genetic material from each parent due to recombination and independent assortment during meiosis, and random fertilisation (Hill & Weir 2011, 2012; Speed & Balding 2015). Even when pedigrees are robust, this study highlights an unavoidable shortcoming as pedigrees do not adequately capture true relatedness between full siblings. We anticipate this uncaptured diversity may prove useful for maximising existing diversity, especially in conservation breeding programmes with relatively few founders (Ballou & Lacy 1995).

Compared to the pedigree-based approach, our empirical data sets (i.e., microsatellites and SNPs) capture more variation between siblings than parents and offspring (Figure 5.2). A broad range of microsatellite-based relatedness estimates were observed in both parent-offspring and sibling relationships, compared to the SNP-based approach. In some instances, even parent-offspring pairings appeared relatively unrelated using microsatellites (e.g., minimum parent-offspring  $R = 0.14$  in kakī and  $R = 0$  in kākārīki karaka), which underscores the lack of precision in this approach and how it could inadvertently lead to poorly informed pairing recommendations. These large ranges of relatedness values using microsatellites can be explained because genetic-based relatedness values between parent-offspring and full siblings are based on allele frequencies, and relatedness between individuals that share common alleles will be substantially lower than individuals that share rare alleles (Wang 2011; Speed & Balding 2015). This bias in relatedness values can be exacerbated when sample sizes are small (Wang 2017), which is typical of conservation breeding programmes. Furthermore, the lack of precision using

microsatellites shown here is consistent with studies that suggest relatively few markers with low allelic diversity are insufficient for estimating relatedness and inbreeding, especially in genetically depauperate species (e.g., Taylor 2015; Taylor *et al.* 2015; Hellmann *et al.* 2016; Escoda *et al.* 2017; Attard *et al.* 2018).

Compared to microsatellite-based relatedness, SNP-based relatedness showed a relatively small range with parent-offspring and full sibling relatedness estimates approximating 0.5, and full siblings showing a wider range of values than parent-offspring relationships (Figure 5.2). Not only is this pattern consistent with expectations given the behaviour of chromosomes during meiosis and random fertilisation, but it also shows more precision than the microsatellite data sets. Other researchers have found similar results in a diverse range of wild taxa, indicating that thousands of genome-wide SNPs show more precision than microsatellites when measuring relatedness and inbreeding (e.g., Hoffman *et al.* 2014; Hellmann *et al.* 2016; Attard *et al.* 2018; Thrasher *et al.* 2018; Lemopoulos *et al.* 2019).

Beyond parent-offspring and full sibling relationships, pedigree and SNP-based relatedness estimates showed the highest concordance with one another among the three approaches used (Figure 5.3). In kakī, the data sets used here include non-captive bred individuals with intensively monitored wild parents. These results provide more credibility to the semi-wild kakī pedigree, where socially monogamous wild pairs of kakī are assumed to be the genetic parents of offspring at nests (but see also Overbeek *et al.* 2017). Still, it should be noted that many pairs with pedigree-based relatedness values of 0 had SNP-based relatedness values ranging upwards of 0.40 in kakī 0.33 and in kākārīki karaka, which approximates first and second order relationships in both species (Figure 5.2). This indicates that pedigree-based *R* between these individuals may be downwardly biased by the

assumption that no variance in relatedness exists amongst founders, missing information, and/or low pedigree depth (Lacy 1995; Balloux *et al.* 2004; Pemberton 2008; Rudnick & Lacy 2008; Tzika *et al.* 2009; Bérénos *et al.* 2014; Kardos *et al.* 2015; Hammerly *et al.* 2016; Hogg *et al.* 2018).

#### **5.2.4.2: Pairing Recommendations**

When these relatedness values are translated into pairing recommendations using MSI scores and MK rank, there is a high concordance between pedigree and SNP-based approaches, with SNP-based MSI scores being significantly higher than pedigree- and microsatellite-based approaches. The latter result is somewhat expected, given that average relatedness estimates using SNPs was highest among the approaches used here, and empirical estimates of relatedness and inbreeding are usually higher than pedigrees as they more effectively capture relatedness between founders or mis-assigned individuals (Hammerly *et al.* 2016; Hogg *et al.* 2018). With that said, when making pairing recommendations using kinship-based pairing decisions (e.g., Ballou & Lacy 1995), it is often the relative kinships between individuals that are more important than absolute values (Galla *et al.* 2019; McLennan *et al.* 2019). This suggests that pedigree and SNP-based approaches both yield similar results for pairing recommendations, with some important differences. For example, while correlation coefficients between these two sets of MSI scores are high relative to other comparisons, there are instances where pairings are considered ‘highly beneficial’ (i.e., MSI category 1) when using the pedigree and ‘very highly detrimental’ (i.e., MSI category 7) when using SNPs. When comparing MK ranks between pedigree- and SNP-based approaches, some kakī and kākāriki karaka experienced large shifts in rank (i.e.,  $\geq 8$  positions difference) depending on the approach used. Although we expect some differences between pedigree- and SNP-based MSI scores and MK ranks, we

attribute these very large differences to errors in the pedigree (e.g., Hammerly *et al.* 2016) or violations of the assumption that there is no variance in founder relationships (e.g., Hogg *et al.* 2018). Of all kakī and kākāriki karaka pairings that experienced a large shift between pedigree- and SNP-based MSI scores, most feature recurring individuals with wild parentage (i.e., founders), and in one recurring occasion, a wild individual (kakī) with high pedigree depth that likely represents an entry error in the pedigree.

#### **5.2.4.3: Management Implications**

Pedigree, genetic, and genomic-based tools each have their advantages to inform conservation management. For example, pedigrees capture both genetic and demographic considerations dating back to the founding of the population, while empirical estimates of relatedness can circumvent pedigree errors and issues surrounding founder relationships by expressing realised relatedness between all individuals. From the results shown here, we recommend that when conservation breeding programmes are poorly pedigreed (i.e., pedigrees of low depth or containing missing data), SNPs should be incorporated to provide a precise indicator of relatedness to genetically inform pairing decisions. The microsatellite panels used here have shown low precision in estimating relatedness, with demonstrated downstream effects for pairing recommendations compared to pedigree- and SNP-based approaches. While more microsatellites could be used to mitigate this shortcoming, other studies indicate that more microsatellites may not equate to higher precision for relatedness estimates and inbreeding coefficients (Robinson *et al.* 2013; Taylor 2015; Taylor *et al.* 2015; Nietlisbach *et al.* 2017). Further, the time and cost associated with building larger microsatellite panels and generating microsatellite data will likely be surpassed by the production of genome-wide SNPs, either by a whole genome resequencing approach as shown in this study, or by a reduced-representation sequencing approaches (e.g., RAD-

sequencing, or Genotyping-by-Sequencing; Narum *et al.* 2013; Galla *et al.* 2016). Currently, for kakī and kākāriki karaka, reduced-representation sequencing is more cost-effective than whole-genome resequencing (i.e., approximately one third of the price, depending on the genome size, as of 2019) — but we foresee more people shifting towards whole genome resequencing in the near future, given the decreasing cost of high throughput sequencing (Hayden 2014) and the ability to ask more research questions using whole genome resequencing data sets (see Future Directions below for details). This is particularly true for birds, whose genomes are small (e.g., 1.05-1.26G) relative to many vertebrates (Zhang *et al.* 2014).

We anticipate SNPs will be particularly applicable in circumstances when pedigrees are the least reliable. For instance, when the founders of a conservation breeding population have no ancestry data available and are likely to be related, SNP-based relatedness estimates between individuals can be used to avoid highly related matings (Hogg *et al.* 2018). This situation may not only coincide with the original founding event of a captive population, but iteratively when individuals are sourced from wild or translocated populations to augment the captive population, as suggested in Frankham (2008) and Hogg *et al.* (2018). For example, in kākāriki karaka, whole genome resequencing has been made available for all current breeding individuals in the conservation breeding programme, including individuals who are founders themselves. Because birds of unknown ancestry are being routinely sourced from highly endangered wild populations, and will also be founders, we anticipate the need for resequencing these birds as they are incorporated into the breeding programme to assess their relatedness to other individuals. In addition to traditional captive bred populations (i.e., *ex situ* management), this approach is applicable to intensively managed wild populations (i.e., *in situ* management), where robust pedigrees

are lacking, but conservation translocations can be informed by relatedness between individuals in a managed landscape (e.g., kākāpō, *Strigops habroptilus*, Elliott *et al.* 2001; scimitar-horned oryx, *Oryx dammah*, Wildt *et al.* 2019).

While we expect SNPs will be important for pairing recommendations moving forward, we do not expect they will eclipse well-established pedigrees, as both approaches have advantages for conservation breeding. Instead, we envision a combined approach where realised relatedness from SNPs can be used to augment data-rich pedigrees. With that said, there are relatively few studies to date that effectively combine existing pedigree data with genomic estimates of relatedness to inform pairing recommendations (but see Hogg *et al.* 2018; Ivy *et al.* 2016). To date, these studies are largely limited to SNPs being used for parentage reconstruction (reviewed in Flanagan & Jones 2019), where unknown or uncertain relationships are reconstructed using empirical data and software (e.g., Whalen *et al.* 2018), and more complete pedigrees are used moving forward. Alternatively, there is an option to produce empirical estimates of relatedness for all founders or breeding individuals in conservation breeding programmes — as suggested in Ivy *et al.* (2016) and practiced in Hogg *et al.* (2018) — and use this baseline of known relatedness moving forward using pedigrees. While the programme PMx allows for the inclusion of empirical data (Lacy *et al.* 2012), this approach requires caution, as the calculation of pedigree-based identity by descent for subsequent generations — including kinship and gene diversity — will be affected by the addition of empirical data (Hogg *et al.* 2018). We acknowledge this approach requires further investigation and validation, particularly for species that receive periodic influx of wild individuals of unknown ancestry in their conservation breeding programme.

#### 5.2.4.4: Future Directions and Concluding Remarks

This study has produced pedigrees and whole genome sequences for two critically-endangered species. Beyond estimating relatedness, these tools provide an exciting opportunity to explore other questions relevant to conservation, such as characterising the genomic-basis of fitness traits, including those associated with inbreeding depression (Kardos *et al.* 2016; but see also Kardos & Shafer 2019) or adaptation to captivity (e.g., Grueber *et al.* 2017). We also envision using the genomic resources developed here to further investigate best practice for making pairing recommendations; for example, agent-based, multi-generational simulations can be used to evaluate whether genome-wide diversity is best maximised using pedigrees, SNPs, or a combination approach.

Given that SNPs have been successfully used to estimate relatedness for different purposes across a wide diversity of taxonomic groups outside of this study (as reviewed in Attard *et al.* 2018), we anticipate a SNP-based approach for estimating relatedness and making subsequent pairing recommendations will be applicable beyond birds. In the meantime, for poorly pedigreed populations worldwide, we recommend a SNP-based approach to estimate relatedness for subsequent pairing recommendations. It should be noted that many approaches used to date have used *de novo* reduced representation approaches (e.g., genotyping-by-sequencing, RADseq; Narum *et al.* 2013) for SNP discovery, which typically have more missing data, lower depth, and fewer SNPs than the reference-guided whole genome resequencing approach used here. While these factors may contribute to bias in relatedness estimates (but see Dodds *et al.* 2015), research still indicates that fewer SNPs, with more missing data and lower depth, than those presented here provide more precision than microsatellites (Attard *et al.* 2018). We expect reduced-representation approaches will persist in the short-term, especially for species with large

and complex genomes (e.g., some fish, amphibians, and invertebrates) that otherwise cannot yet be affordably resequenced across entire conservation breeding programmes. With that said, we also expect whole-genome resequencing projects like ours will gain momentum in the years to come, as these data can be better leveraged to address multiple questions related to conservation genetic management (Harrison *et al.* 2014; see also above). In the meantime, we look forward to seeing more poorly pedigreed conservation breeding programmes for taxonomically diverse species from around the world incorporate SNPs for estimating relatedness to inform pairing decisions.

### 5.2.5: Acknowledgements

We are grateful for the support of Te Rūnanga o Ngāi Tahu, Te Ngāi Tūāhuriri Rūnanga, Te Rūnanga o Arowhenua, Te Rūnanga o Waihao and Te Rūnanga o Moeraki. We thank all members of the Kakī Recovery Group and the Orange-fronted Parakeet Recovery Group for their assistance with sample collection and their ongoing support. This research has been funded by the Ministry of Business, Innovation and Employment (MBIE) Endeavour Fund (UOCX1602 awarded to TES), the Brian Mason Scientific and Technical Trust (awarded to SJG and TES), and the Mohua Charitable Trust (awarded to TES).

### 5.2.6: References

- Andrews BJ (2013) *Conservation genetics of the orange-fronted kākāriki (Cyanoramphus malherbi)* (Masters Thesis). University of Canterbury.
- Andrews BJ, Hale ML, Steeves TE (2013) Characterisation of microsatellite loci in the critically endangered orange-fronted kākāriki (*Cyanoramphus malherbi*) isolated using genomic next generation sequencing. *Conservation Genetics Resources*, **5**(1), 235–237.
- Andrews S (2010) *FastQC: a quality control tool for high throughput sequence data* [Computer Software]. Retrieved from <https://github.com/s-andrews/FastQC>.
- Attard CRM, Beheregaray LB, Möller LM (2018) Genotyping-by-sequencing for estimating relatedness in nonmodel organisms: Avoiding the trap of precise bias. *Molecular Ecology Resources*, **18**(3), 381–390.



- Attard CRM, Möller, LM, Sasaki M, *et al.* (2016) A novel holistic framework for genetic-based captive-breeding and reintroduction programs: Holistic Framework for reintroductions. *Conservation Biology*, **30(5)**, 1060–1069.
- Ballou, JD (1983) Calculating inbreeding coefficients from pedigrees. In *Genetics and Conservation*. Benjamin/Cummings Publishing Co. 509–516.
- Ballou JD, Lacy RC (1995). Identifying genetically important individuals for management of genetic variation in pedigreed populations. In *Population Management for Survival and Recovery*. Columbia University press. 78–111.
- Ballou JD, Lees C, Faust LJ, Long S, Lynch C, Bingaman LL, Foose TJ (2010) Demographic and genetic management of captive populations. In *Wild Mammals in Captivity: Principles and Techniques for Zoo Management*. University of Chicago Press. 219-252.
- Balloux F, Amos W, Coulson T (2004) Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, **13(10)**, 3021–3031.
- Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2014). Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. *Molecular Ecology*, **23(14)**, 3434–3451.
- Bergner LM, Jamieson IG, Robertson BC (2014). Combining genetic data to identify relatedness among founders in a genetically depauperate parrot, the kakapo (*Strigops habroptilus*). *Conservation Genetics*, **15(5)**, 1013–1020.
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30(15)**, 2114–2120.
- Chan C, Ballantyne KN, Lambert DM, Chambers GK (2005) Characterization of variable microsatellite loci in Forbes' parakeet (*Cyanoramphus forbesi*) and their use in other parrots. *Conservation Genetics*, **6(4)**, 651–654.
- Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS (2011) Meraculous: de novo genome assembly with short paired-end reads. *PloS One*, **6(8)**, e23501.
- Charlesworth D, Charlesworth B (1987) Inbreeding depression and its evolutionary consequences. *Annual Review of Ecology and Systematics*, **18**, 237–268.
- Crow JF, Kimura M (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row, Publishers.
- Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27(15)**, 2156–2158.

- De Fraga R, Lima AP, Magnusson WE, Ferrão M, Stow AJ (2017) Contrasting patterns of gene flow for Amazonian snakes that actively forage and those that wait in ambush. *Journal of Heredity*, **108**(5), 524–534.
- de Villemereuil P, Rutschmann A, Lee KD, Ewen JG, Brekke P, Santure AW (2019) Little adaptive potential in a threatened passerine bird. *Current Biology*, **29**(5), 889–894. e3.
- Dodds KG, McEwan JC, Brauning R, Anderson RM, van Stijn TC, Kristjánsson T, Clarke SM (2015). Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, **16**(1), 1047.
- Ellegren H (2013). The evolutionary genomics of birds. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 239–259.
- Elliott GP, Merton DV, Jansen PW (2001). Intensive management of a critically endangered species: the kakapo. *Biological Conservation*, **99**(1), 121–133.
- Escoda L, González-Esteban J, Gómez A, Castresana J (2017). Using relatedness networks to infer contemporary dispersal: Application to the endangered mammal *Galemys pyrenaicus*. *Molecular Ecology*, **26**(13), 3343–3357.
- Faust LJ, Bergström YM, Thompson SD, Bier L (2018) *PopLink Version 2.5* [Computer Software]. Retrieved from: <https://www.lpzoo.org/conservation-science/projects/poplink>.
- Forstmeier W, Schielzeth H, Mueller JC, Ellegren H, Kempenaers B (2012) Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. *Molecular Ecology*, **21**(13), 3237–3249.
- Frankham R (2008) Genetic adaptation to captivity in species conservation programs. *Molecular Ecology*, **17**(1), 325–333.
- Fridolfsson AK, Ellegren H (1999) A simple and universal method for molecular sexing of non-ratite birds. *Journal of Avian Biology*, 116–121.
- Galla SJ, Buckley TR, Elshire R, *et al.* (2016) Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances. *Molecular Ecology*, **25**(21), 5267–5281.
- Galla SJ, Forsdick NJ, Brown L, *et al.* (2019) Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management. *Genes*, **10**(1), 9.
- Gooley R, Hogg CJ, Belov K, Grueber CE (2017) No evidence of inbreeding depression in a Tasmanian devil insurance population despite significant variation in inbreeding. *Scientific Reports*, **7**, 1–11.

- Griffiths R, Double MC, Orr K, Dawson RJ (1998) A DNA test to sex most birds. *Molecular Ecology*, **7(8)**, 1071–1075.
- Grueber CE, Reid-Wainscoat EE, Fox S, Belov K, Shier DM, Hogg CJ, Pemberton D (2017) Increasing generations in captivity is associated with increased vulnerability of Tasmanian devils to vehicle strike following release to the wild. *Scientific Reports*, **7**, 1-7.
- Grueber CE, Fox S, McLennan EA, Gooley RM, Pemberton D, Hogg CJ, Belov K (2019) Complex problems need detailed solutions: Harnessing multiple data types to inform genetic management in the wild. *Evolutionary Applications*, **12(2)**, 280–291.
- Hagen EN, Hale ML, Maloney RF, Steeves TE (2011) Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the Black Stilt *Himantopus novaezelandiae*. *Ibis*, **153(3)**, 556–561.
- Hammerly SC, de la Cerda DA, Bailey H, Johnson JA (2016) A pedigree gone bad: increased offspring survival after using DNA-based relatedness to minimize inbreeding in a captive population. *Animal Conservation*, **19(3)**, 296–303.
- Hammerly SC, Morrow ME, Johnson JA (2013) A comparison of pedigree- and DNA-based measures for identifying inbreeding depression in the critically endangered Attwater's Prairie-chicken. *Molecular Ecology*, **22(21)**, 5313–5328.
- Hayden EC (2014) The \$1,000 genome. *Nature*, **507**, 295–295.
- He X, Johansson ML, Heath DD (2016) Role of genomics and transcriptomics in selection of reintroduction source populations. *Conservation Biology*, **30(5)**, 1010–1018.
- Hellmann JK, Sovic MG, Gibbs HL, *et al.* (2016) Within-group relatedness is correlated with colony-level social structure and reproductive sharing in a social fish. *Molecular Ecology*, **25(16)**, 4001–4013.
- Hedrick PW, Lacy RC (2015) Measuring relatedness between inbred individuals. *Journal of Heredity*, **106**, 20-25.
- Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, **93(1)**, 47–64.
- Hill WG, Weir BS (2012) Variation in actual relationship among descendants of inbred individuals. *Genetics Research*, **94(05)**, 267–274.
- Hoffman JI, Simpson F, David P, *et al.* (2014). High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences USA*, **111(10)**, 3775–3780.

- Hogg CJ, Wright B, Morris KM, Lee AV, Ivy J. A., Grueber, C. E., & Belov, K. (2018). Founder relationships and conservation management: empirical kinships reveal the effect on breeding programmes when founders are assumed to be unrelated. *Animal Conservation*, **22**(4), 348-361.
- Ivy JA, Lacy RC (2012) A comparison of strategies for selecting breeding pairs to maximize genetic diversity retention in managed populations. *Journal of Heredity*, **103**(2), 186–196.
- Ivy JA, Miller A, Lacy RC, DeWoody JA (2009) Methods and prospects for using molecular data in captive breeding programs: An empirical example using parma wallabies (*Macropus parma*). *Journal of Heredity*, **100**(4), 441–454.
- Ivy JA, Putnam AS, Navarro AY, Gurr J, Ryder OA (2016) Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, **107**(5), 403–412.
- Jiménez-Mena B, Schad K, Hanna N, Lacy RC (2016) Pedigree analysis for the genetic management of group-living species. *Ecology and Evolution*, **6**(10), 3067–3078.
- Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**(10), 2511–2523.
- Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, **115**(1), 63–72.
- Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW (2016) Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, **9**, 1205-1218.
- Kardos M, Shafer ABA (2019) The peril of gene-targeted conservation. *Trends in Ecology and Evolution*, **33**(11), 827-839.
- Kearvell, JC, Legault AJ (2017) Orange-fronted parakeet. In *New Zealand Birds Online*. Retrieved from [www.nzbirdsonline.org.nz](http://www.nzbirdsonline.org.nz).
- Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics*, **189**(1), 237–249.
- Kielbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Research*, **21**(3), 487–493.
- Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, Forstmeier W (2015) Quantifying realized inbreeding in wild and captive animal populations. *Heredity*, **114**(4), 397–403.

- Knief U, Kempenaers B, Forstmeier W (2017) Meiotic recombination shapes precision of pedigree- and marker-based estimates of inbreeding. *Heredity*, **118**, 239–248.
- Krueger F (2019) *Trim Galore* [Computer Software]. Retrieved from <https://github.com/FelixKrueger/TrimGalore>.
- Lacy RC (1995) Clarification of genetic terms and their use in the management of captive populations. *Zoo Biology*, **14(6)**, 565–577.
- Lacy RC (2012) Extending pedigree analysis for uncertain parentage and diverse breeding systems. *Journal of Heredity*, **103(2)**, 197–205.
- Lacy RC, Ballou JD, Pollak JP (2012) PMx: software package for demographic and genetic analysis and management of pedigreed populations. *Methods in Ecology and Evolution*, **3**, 433–437.
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9(4)**, 357–359.
- Lemopoulos A, Prokkola JM, Uusi-Heikkilä S, *et al.* (2019) Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution*, **9(4)**, 2106–2120.
- Li H, Handsaker B, Wysoker A, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25(16)**, 2078–2079.
- Li MH, Strandén I, Tiirikka T, Sevón-Aimonen ML, Kantanen J (2011) A Comparison of approaches to estimate the inbreeding coefficient and pairwise relatedness using genomic and pedigree data in a sheep population. *PLoS ONE*, **6(11)**, e26256.
- Li N (2010) *identity: Jacquard condensed coefficients of identity* [R Programme]. Retrieved from <https://cran.r-project.org/web/packages/identity/index.html>
- Long S, Dorsey C, & Boyle P (2011) Status of Association of Zoos and Aquariums cooperatively managed populations. *WAZA Magazine*, **12**, 15–18.
- Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27(6)**, 764–770.
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, **17(1)**, 10–12.
- McLennan EA, Wright BR, Belov K, Hogg CJ, Grueber CE (2019) Too much of a good thing? Finding the most informative genetic data set to answer conservation questions. *Molecular Ecology Resources*, **19(3)**, 659–671.

- Mitchell AA, Lau J, Chemnick LG, Thompson EA, Alberts AC, Ryder OA, Gerber GP (2011) Using microsatellite diversity in wild Anegada iguanas (*Cyclura pinguis*) to establish relatedness in a captive breeding group of this critically endangered species. *Conservation Genetics*, **12**(3), 771–781.
- Moore CT, Converse SJ, Folk MJ, Runge MC, Nesbitt SA (2012) Evaluating release alternatives for a long-lived bird species under uncertainty about long-term demographic rates. *Journal of Ornithology*, **152**(2), 339–353.
- Moraga R (2018a) *concatenate\_genome.pl* [Custom Perl Script]. Retrieved from [https://github.com/sgalla32/concatenate\\_genome](https://github.com/sgalla32/concatenate_genome).
- Moraga R (2018b) *split\_bamfile\_tasks.pl* [Custom Perl Script]. Retrieved from [https://github.com/Lanilen/SubSampler\\_SNPcaller](https://github.com/Lanilen/SubSampler_SNPcaller).
- Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**(11), 2841–2847.
- Nietlisbach P, Keller LF, Camenisch G, Guillaume F, Arcese P, Reid JM, Postma E (2017) Pedigree-based inbreeding coefficient explains more variation in fitness than heterozygosity at 160 microsatellites in a wild bird population.
- Overbeek AL, Hauber ME, Brown E, Cleland S, Maloney RF, Steeves TE (2017) Evidence for brood parasitism in a critically endangered Charadriiform with implications for conservation. *Journal of Ornithology*, **158**(1), 333–337.
- Pierce RJ (2013) Black stilt. In Miskelly, C.M. (ed.) *New Zealand Birds Online*. [www.nzbirdsonline.org.nz](http://www.nzbirdsonline.org.nz).
- Premachandra HKA, Nguyen NH, Knibb W (2019) Effectiveness of SNPs for parentage and sibship assessment in polygamous yellowtail kingfish *Seriola lalandi*. *Aquaculture*, **499**, 24–31.
- Peakall ROD, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**(1), 288–295.
- Pemberton J (2004) Measuring inbreeding depression in the wild: the old ways are the best. *Trends in Ecology & Evolution*, **19**(12), 613–615.
- Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society B: Biological Sciences*, **275**(1635), 613–621.
- Reed CEM (1998) *Management plan for captive black stilts*. Biodiversity Recovery Unit, Department of Conservation.

- Robertson BC, Gemmell NJ (2006) PCR-based sexing in conservation biology: Wrong answers from an accurate methodology? *Conservation Genetics*, **7**(2), 267–271.
- Robinson SP, Simmons LW, Kennington WJ (2013) Estimating relatedness and inbreeding using molecular markers and pedigrees: the effect of demographic history. *Molecular Ecology*, **22**(23), 5779–5792.
- Rudnick JA, Lacy RC (2008) The impact of assumptions about founder relationships on the effectiveness of captive breeding strategies. *Conservation Genetics*, **9**(6), 1439–1450.
- Sanders MD, Maloney RF (2002) Causes of mortality at nests of ground-nesting birds in the Upper Waitaki Basin, South Island, New Zealand: a 5-year video study. *Biological Conservation*, **106**(2), 225–236.
- Santure AW, Stapley J, Ball AD, Birkhead TR, Burke T, Slate J (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**(7), 1439–1451.
- Slate J, David P, Dodds KG, Veenvliet BA, Glass BC, Broad TE, McEwan JC (2004) Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. *Heredity*, **93**(3), 255–265.
- Slotta-Bachmayr L, Boegel R, Kaczensky P, Stauffer C, Walzer C (2004) Use of population viability analysis to identify management priorities and success in reintroducing Przewalski's horses to southwestern Mongolia. *The Journal of Wildlife Management*, **68**(4), 790–798.
- Smilkov D, Thorat N, Nicholson C, Reif E, Viégas FB, Wattenberg M (2016) Embedding Projector: Interactive Visualization and Interpretation of Embeddings. *ArXiv*, [abs/1611.05469](https://arxiv.org/abs/1611.05469).
- Smith DHV, Converse SJ, Gibson KW, Moehrensclager A, Link WA, Olsen GH, Maguire K (2011) Decision analysis for conservation breeding: Maximizing production for reintroduction of whooping cranes. *The Journal of Wildlife Management*, **75**(3), 501–508.
- Smouse RPP, Peakall R (2012) GenALEX 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, **28**(19), 2537–2539.
- Soulé ME (1985) What is conservation biology? *BioScience*, **35**(11), 727–734.
- Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, **16**(1), 33–44.

- Steeves TE, Hale ML, Gemmell NJ (2008) Development of polymorphic microsatellite markers for the New Zealand black stilt (*Himantopus novaezelandiae*) and cross-amplification in the pied stilt (*Himantopus himantopus leucocephalus*). *Molecular Ecology Resources*, **8**(5), 1105–1107.
- Steeves TE, Maloney RF, Hale ML, Tylianakis JM, Gemmell NJ (2010) Genetic analyses reveal hybridization but no hybrid swarm in one of the world's rarest birds. *Molecular Ecology*, **19**(23), 5090–5100.
- Sutton JT, Robertson BC, Jamieson IG (2011) Dye shift: a neglected source of genotyping error in molecular ecology. *Molecular Ecology Resources*, **11**(3), 514–520.
- Taylor HR (2015) The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and Evolution*, **5**(15), 3140–3150.
- Taylor HR, Kardos MD, Ramstad KM, Allendorf FW (2015) Valid estimates of individual inbreeding coefficients from marker-based pedigrees are not feasible in wild populations with low allelic diversity. *Conservation Genetics*, **16**(4), 901–913.
- Tenhumberg B, Tyre AJ, Shea K, Possingham HP (2004) Linking wild and captive populations to maximize species persistence: optimal translocation strategies. *Conservation Biology*, **18**(5), 1304–1314.
- Tzika AC, Remy C, Gibson R, Milinkovitch MC (2009) Molecular genetic analysis of a captive-breeding program: the vulnerable endemic Jamaican yellow boa. *Conservation Genetics*, **10**(1), 69–77.
- Visscher PM, Medland SE, Ferreira MAR, *et al.* (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, **2**(3), e41.
- Walters JR, Derrickson SR, Michael FD, Haig SM, Marzluff JM, Wunderle JM (2010) Status of the California condor (*Gymnogyps californianus*) and efforts to achieve its recovery. *The Auk*, **127**(4), 969–1001.
- Waples RK, Albrechtsen A, Moltke I (2019) Allele-frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Molecular Ecology*, **28**, 35–48.
- Wang J (2007) Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetics Research*, **89**(3), 135–153.
- Wang J (2011) coancestry: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, **11**(1), 141–145.
- Wang J (2016) Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, **107**, 4–13.



- Wang J (2017) Estimating pairwise relatedness in a small sample of individuals. *Heredity*, **119**(5), 302.
- Whalen A, Gorjanc G, Hickey JM (2018) Parentage assignment with genotyping-by-sequencing data. *Journal of Animal Breeding and Genetics*, **136**(2), 102-112.
- Wildt D, Miller P, Koepfli KP, *et al.* (2019) Breeding centers, private ranches, and genomics for creating sustainable wildlife populations. *BioScience*, 69(11), 928-943.
- Willoughby JR, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, DeWoody JA (2015) The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology*, **24**(1), 98–110.
- Zhang G, Li B, Li C, Gilbert MTP, Jarvis ED, Wang J (2014) Comparative genomic data of the Avian Phylogenomics Project. *GigaScience*, **3**(1), 26.
- Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. *Bioinformatics*, **29**(21), 2669–2677.

#### 5.2.8: Data accessibility

Genomic data provided in this manuscript are available through a password protected server on the Conservation, Systematics and Evolution Research Team’s website (<http://www.uconsert.org/data/>). Kakī and kākāriki karaka are taonga (treasured) species. For Māori (the indigenous people of Aotearoa), all genomic data obtained from taonga species have whakapapa (genealogy that includes people, plants and animals, mountains, rivers and winds) and are therefore taonga in their own right. Thus, these data are tapu (sacred) and tikanga (customary practices, protocols, and ethics) determine how people interact with it. To this end, the passwords for the genomic data in this manuscript will be made available to researchers on the recommendation of the kaitiaki (guardians) for the iwi (tribes) that affiliate with kakī and kākāriki karaka.

## Chapter 6: General Discussion and Future Directions

### 6.1: Brief Synopsis

This thesis generates and examines tools for making pairing recommendations in conservation breeding programmes and serves as an exemplar for how research teams might transition from using genetic to genomic technologies in conservation. Pedigrees have traditionally been used to pair individuals in conservation breeding programmes in an effort to minimise inbreeding, maximise diversity, and reduce extinction risk (Ballou & Lacy 1995; Lacy 2012). However, as highlighted in **Chapters 2 and 5**, the utility of pedigrees may be hampered by the assumption that all founders are unrelated, including individuals of unknown parentage that may be periodically sourced into conservation breeding programmes (Ivy *et al.* 2016; Hogg *et al.* 2018). While genetic markers (i.e., microsatellites) can be used to produce empirical estimates of relatedness in lieu of robust pedigrees, research in **Chapter 5** indicates they are insufficient for estimating relatedness in kakī and kākāriki karaka, and more precise measures are obtained from thousands of genome-wide SNPs. SNP-based relatedness estimates and pairing recommendations also show a high degree of concordance with one another, indicating that SNP-based estimates of relatedness will be the best tool of choice for pairing recommendations in poorly-pedigreed conservation breeding programmes. While SNPs may provide effective and efficient estimates of relatedness, many researchers are still transitioning towards genomic technologies. **Chapters 3 and 4** provide approaches to expedite the transition including partnering with primary industry researchers and using readily available reference genomes from closely-related species for marker discovery.

## 6.2: The Conservation Genomics Gap, in Retrospect

This thesis represents the first endeavour in our research team to apply genomic technologies to conservation management. This transition was facilitated by early partnerships with primary industry researchers (**Chapter 3**, Galla *et al.* 2016), which allowed for rapid upskilling in HTS data storage and curation (e.g., workshops by John McCallum, Plant & Food Research), read pre-processing (e.g., workshops by Roger Moraga, AgResearch/Tea Break Bioinformatics, Ltd.), SNP discovery (e.g., workshops by Roger Moraga), and the production of relatedness and diversity estimates (as per the KGD method, developed for primary industry species; Dodds *et al.* 2015). These lessons have enabled the conservation research presented in **Chapters 4 and 5**, and as envisioned in Galla *et al.* 2016 (**Chapter 3**), partnerships between researchers in primary industry and conservation continue to lead to mutually beneficial outcomes, including co-developed research in our research team (e.g., <http://www.biologicalheritage.nz/news/news/cutting-edge-tech>) and collaborative research on a national level (e.g., Genomics Aotearoa, a multi-disciplinary genomics platform focusing on primary production, human health, and the environment <https://www.genomics-aotearoa.org.nz/projects>).

**Chapter 4** addresses another component of the conservation genomics gap: the availability of genomic resources for threatened species. My co-authors and I determine that readily available and closely-related reference genomes can be used for SNP discovery to generate diversity estimates for use in conservation. With genomes representing all known bird families becoming available in the relatively near future as a part of Birds 10,000 Genomes Project (Zhang *et al.* 2014; also see <https://b10k.genomics.cn/progress.html>), we expect these genomes will be a valuable resource for those who do not have readily available references for marker discovery for use in conservation. It should be noted that

the quality of these short-read *de novo* genomes should be assessed prior to use, as these genomes may have substantial assembly gaps that can affect SNP discovery (Peona *et al.* 2018; Galla & Forsdick *et al.* 2019). While the use of closely related reference genomes appears particularly successful for birds who have small and relatively conserved genomes (Zhang *et al.* 2014), this approach will need to be evaluated in other taxonomic groups with larger, more complex, and differentiated genomes, including mammals, reptiles, amphibians, fish, and invertebrates. With more genomic resources becoming available through the Genome 10K/Vertebrate Genomes Project (G10K; Genome 10K Scientists 2009), the 5,000 Insect Genome Project (i5K; Robinson *et al.* 2011), the 1,000 Plants Project (1KP; Matasci *et al.* 2014), the Oz Mammalian Genomics (Duchene *et al.* 2018), and the Earth BioGenome Project (Lewin *et al.* 2018), there will be opportunities for this approach be evaluated in other taxonomic groups.

### 6.3: Pairing Recommendations

**Chapter 5** provides strong evidence against the use of microsatellite markers for estimating relatedness in conservation breeding programmes, with more precision found with genome-wide SNPs. This research corroborates other studies that suggest microsatellites provide poor resolution for relatedness estimates (e.g., Escoda *et al.* 2016; Hellman *et al.* 2016; Attard *et al.* 2018). Pedigree-based relatedness estimates and downstream pairing recommendations were most similar between pedigree- and SNP-based approaches, indicating that SNPs are an appropriate tool for making pairing recommendations in the absence of robust pedigrees. While this has broad implications for poorly pedigreed conservation breeding programmes around the world, these results also have direct implications for kakī and kākāriki karaka management.

As discussed in **Chapters 2** and **5**, empirical estimates of relatedness can be used to inform the kakī pedigree. Because this pedigree is still relatively shallow and historically includes many individuals of unknown parentage (**Chapter 2**), SNP-based relatedness can be used to augment the pedigree and allow for more accurate estimates of kinship and inbreeding that are not biased by the assumption that founders are unrelated (Russello & Amato 2004). Genetic and/or genomic data can also be used to resolve individuals of unknown parentage as they arise in the pedigree. For example, during the 2017-2018 breeding season kakī had a record number of breeding pairs in the wild. For the first time since the initiation of the kakī recovery programme, one wild pair successfully raised a full clutch of chicks to fledging age (Department of Conservation, *personal comm.*; Figure 6.1). Because these offspring fledged before being colour banded, they have no identification associating them as offspring of this wild pair. If these birds survive to reproductive age, the Department of Conservation can trap, band, and sample these birds and genetics and/or genomics can be used to determine whether they are indeed offspring of this pair to inform the pedigree.

More unknown individuals will likely become a part of the population as kakī numbers continue to rise, especially with the advent of Te Manahuna Aoraki (<https://www.temanahunaoraki.org/>). This project aims to use a combination of predator fencing, natural barriers, and vast predator trapping networks to manage habitat in Te Manahuna/The Mackenzie Basin. If the wild population experiences significant growth, the Kakī Recovery Group may retire the captive population and focus on the wild recovery efforts, with less priority given to individual-based management (Department of Conservation, *personal comm.*). In the short- to medium-term, we expect that empirical estimates of relatedness will be helpful for pedigree management. We also anticipate that

learnings using kakī as a Proof-of-Concept, as shown in **Chapters 2, 4, and 5**, will be readily applied to other species with conservation breeding programmes.



*Figure 6. 1: A wild pair of kakī and their full clutch of fledged offspring, found in Te Manahuna during the 2017-2018 breeding season. Photo used with permission, courtesy of Jemma Welch, Department of Conservation.*

For kākāriki karaka, genome-wide SNPs will be used in the near future to incorporate individuals of unknown parentage into the kākāriki karaka captive population. As described in **Chapter 2**, the Kākāriki Karaka Recovery Group has recently prioritised the formation of additional captive pairs. To this end, eggs have been harvested from the nests of two different pairs from Oruawairua/Blumine Island, a former captive release site, for inclusion into the captive breeding programme. Captive females of known parentage were also released into the Poulter Valley of North Canterbury to entice the few remaining wild males into conspecific pairings. This approach was successful for one pair, and eggs from this pair have already been harvested and reared in captivity, with chicks being incorporated into the conservation breeding programme (Department of Conservation, *personal comm.*). While whole genome resequencing is now available for all current captive breeding kākāriki karaka, resequenced genomes are also being produced for all individuals sourced from

Oruawairua and the semi-unknown Poulter Valley pair. To avoid the downward bias experienced when founders are treated as unrelated (Russello & Amato 2004; Ivy *et al.* 2016; Hogg *et al.* 2018), we aim to estimate SNP-based relatedness between these unknown birds to all others in the captive population to avoid pairing of individuals that are inadvertently related to one another.

Moving forward, we anticipate that a combination of pedigree- and SNP-based relatedness will be an effective approach for integrating individuals of unknown parentage into the pedigree. However, how to best combine SNP-based estimates of relatedness with existing pedigree resources for both kakī and kākāriki karaka breeding programmes is still under consideration. As described in **Chapter 5**, there is potential for creating SNP-based estimates of relatedness for all current breeding individuals in a conservation breeding programme, with pairing recommendations made based on empirical relatedness between individuals, as described in Ivy *et al.* (2016). These empirical estimates of relatedness can be incorporated into pedigree software such as PMx (Lacy *et al.* 2012) to make pairing recommendations using empirical data alone. Beyond this approach, it is possible for empirical estimates of relatedness to be used as a baseline, with pedigree-based estimates being produced in subsequent generations. However, as discussed in Hogg *et al.* (2018), downstream calculations of gene diversity and kinship may be affected by the combination of identity-by-state (i.e., empirical) and identity-by descent (pedigree) values. Therefore, studies are warranted to test the effectiveness of this approach and how it may affect diversity estimates after several generations.

As discussed in **Chapter 5**, beyond the biases from individuals of unknown parentage, there are other differences between pedigree- and SNP-based approaches that are worth consideration. For example, pedigrees are based on the probability of Mendelian

inheritance, and as such, do not take into account the different genetic information that siblings inherit as a result of meiosis and random fertilisation (Visscher *et al.* 2006; Hill & Weir 2011, 2012; Speed & Balding 2015). Because pedigrees do not capture this diversity, and there is so little diversity remaining in critically endangered populations, we hypothesise that SNP-based estimates of relatedness may eclipse pedigree-based relatedness estimates for making pairing recommendations that maximise diversity in populations, as suggested in Kardos *et al.* (2015) and Wang (2016). Given the reference genomes, whole genome resequencing data, and pedigrees that are available for kakī and kākāriki karaka, there is an opportunity to test which approach results in higher genome-wide diversity using the data sets already produced in this thesis (Figure 6.2). Using the 36 kakī and 36 kākāriki karaka studied in **Chapters 4 and 5**, we envision making pairing recommendations in each species that represent the ‘best’ pairs as defined using the mate suitability index in PMx (Lacy *et al.* 2012; see also **Chapter 5**) using pedigree-, SNP-, and SNP-informed pedigree approaches for estimating kinship or relatedness over several generations. The pedigree- and SNP-based approaches have described previously (**Chapter 5**). The SNP-informed pedigree approach will use only SNP-based relatedness values in the first generation of the simulation, and pedigree-based relatedness in subsequent generations. Using the available resequenced genomes for these pairs, we will simulate meiosis and random fertilisation between these genomes to create ‘offspring’ that would result from pairings specified from pedigree-, SNP-, and SNP-informed pedigree approaches. This simulation will initially be repeated for five generations (i.e., relatively few generations, to capture breeding strategies used to avoid inbreeding and adaptation to captivity; Frankham 2008; Frankham *et al.* 2014) and SNPs from the final simulated generation will be compared to determine which approach best maximises empirical genome-wide diversity.



Because there are pedigree and genomic resources for the endangered tuturuatu/shore plover (*Thinornis novaeseelandiae*) being developed in our research team, there is an opportunity to extend this simulation to also include this species.

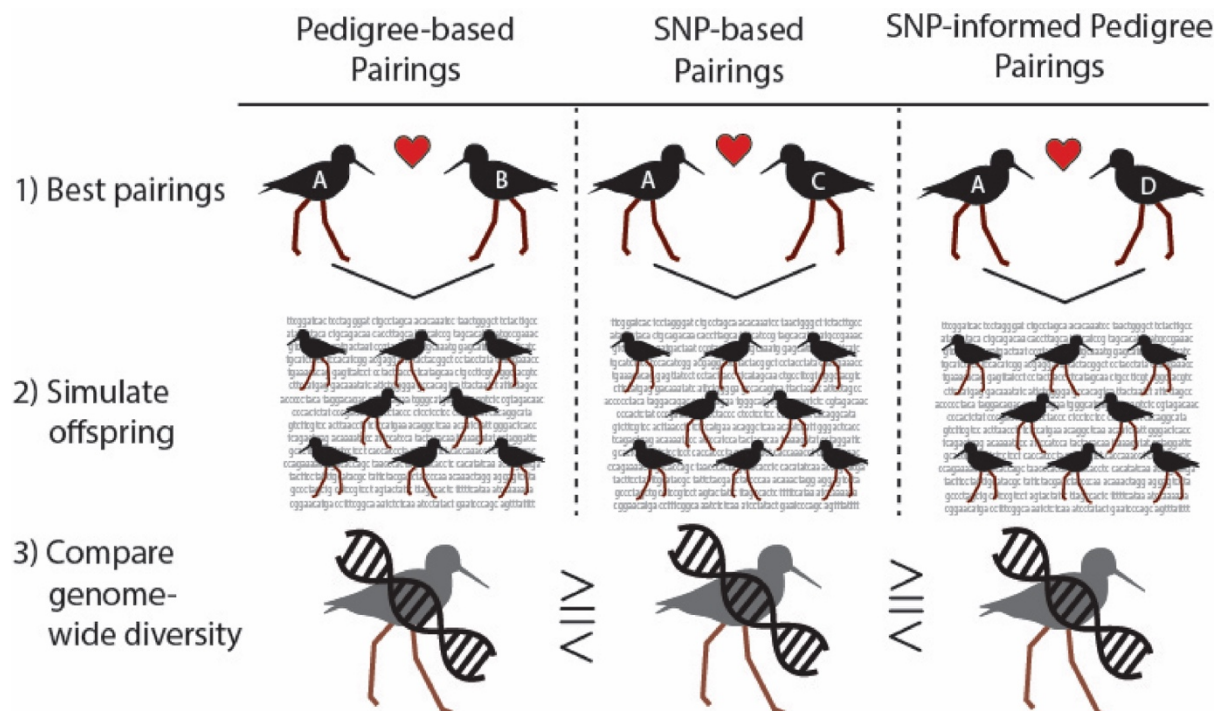


Figure 6. 2: Schematic of suggested simulation study to determine whether pedigree-, SNP-, or SNP-informed pedigree pairings will result in more genome-wide diversity after many generations.

## 6.4: Pedigrees in the Genomics Era

During the course of this thesis, over 500 hours were dedicated towards building and validating the kakī pedigree, with additional work required for both the kakī and kākāriki karaka pedigrees in the future as more individuals are added and errors continue to be resolved. Some have questioned the utility of pedigrees for estimating relatedness in the genomics era (Speed & Balding 2015), as genome-wide SNPs could be used as a precise and time-efficient approach that circumvents pedigree issues related to data entry errors (e.g., Hammerly et al. 2016) and the assumption of no variance in founder relationships (e.g.,

Hogg et al. 2018). In the genomics era, I contend that pedigrees are still valuable in their own right, with or without augmentation from empirical estimates of relatedness.

The kakī and kākāriki karaka pedigrees presented here — albeit with missing information and/or low depth (**Chapters 2 and 5**) — represent hundreds to thousands of individuals and their life history information. At this point in time, it would not be cost effective to resequence whole genomes for all individuals that are captured by a pedigree, nor are there tissue samples available from the founders of many breeding programmes, including those for kakī and kākāriki karaka. Therefore, pedigrees document ancestry amongst individuals that may not be possible to capture using empirical estimates of relatedness. In addition to ancestry information, pedigrees have invaluable life history data that can be mined to elucidate trait heritability (e.g., de Villemereuil et al. 2019), potential inbreeding depression (e.g., Grueber et al. 2010), or factors that are influencing wild and captive survivorship. Beyond the utility of pedigrees for future research, the kakī and kākāriki karaka pedigrees have been a catalyst for forging and solidifying relationships with conservation practitioners. For example, resolving kakī and kākāriki karaka pedigree errors was a team effort between geneticists and practitioners, which resulted in dialogue between both parties and the development of mutual understanding and trust. In a New Zealand context, we also expect that pedigrees can serve to build relationships with mana whenua, as pedigrees are a visual representation of whakapapa (i.e., genealogy) that is central to Te Ao Māori (i.e., the Māori world view; Collier-Robinson et al. 2019).

As discussed in the previous section, we do not anticipate that genome-wide SNPs will eclipse pedigrees. Instead, we envision that the strengths of empirical-based estimates of relatedness can be used to augment pedigrees for kakī, kākāriki karaka, and other pedigreed species worldwide.

## 6.5: Conservation Genomics, Moving Forward

Not only does this thesis provide a better understanding for relatedness and pairing decisions for kakī and kākārīki karaka, it also has provided rich research resources for both species, including pedigrees, *de novo* reference genomes, and 36+ resequenced genomes (**Chapters 2, 4, 5**). In addition to the existing kakī (short-read) reference genome, a high-quality reference genome will soon be generated for kakī as part of Genomics Aotearoa's High Quality Genomes project (<https://www.genomics-aotearoa.org.nz/projects/high-quality-genomes>). To produce this high quality genome, long-read sequencing technologies will be used to augment the existing short-read genome by producing longer scaffolds to close gaps (Rhoads & Au 2015), transcriptome data will be used to annotate the genome with species-specific gene expression (He *et al.* 2016), and Hi-C 3 dimensional genomic analyses (van Berkum *et al.* 2010) will be used to chromosomally assemble the genome, along with karyotyping for validation. Beyond asking questions regarding neutral genome-wide diversity (this thesis), it is also possible to address targeted questions regarding functional variation in kakī (Harrison *et al.* 2014).

Exploring the genomic basis of fitness-related traits is especially apposite for threatened species with maladaptive traits that may hamper species recovery (Kardos *et al.* 2016a, but see also Kardos & Shafer 2019). For example, researchers are actively exploring the genomic basis of chondrodystrophy, an autosomally inherited and lethal form of dwarfism found in California condors (*Gymnogyps californianus*), to inform captive conservation strategies (Ralls *et al.* 2000; Romanov *et al.* 2006; Ryder *et al.* 2016; Cynthia Steiner, *personal comm.*). There have also been efforts to uncover genomic regions under selection for Tasmanian devils (*Sarcophilus harrisii*) in response to facial tumour disease, a lethal and transmissible cancer that has devastated the population (Lazenby *et al.* 2018). To

date, researchers have used allele frequency differences amongst populations of devils throughout time to detect genomic regions under selection in response to facial tumour disease (Epstein *et al.* 2016; Hubert *et al.* 2018) and regions that may underpin how devils survive the disease (Wright *et al.* 2017), which is an exciting development given the relatively low sample sizes used. Given the available resources for kakī, including 15+ years of complete tissue sampling, the pedigree resources developed here (**Chapters 2 and 5**), the short-read reference genome (**Chapter 5**), 36 resequenced genomes (**Chapters 4 and 5**), and the upcoming slated high quality reference genome (i.e., gene-annotated and chromosomally-assembled), there is an opportunity to explore the genomic basis of fitness-related traits of interest in this species.

For example, a 10 year study by Hagen *et al.* (2011) demonstrates a significant negative relationship between kakī breeding pair relatedness and hatching success, indicating that inbreeding may result in reduced reproductive fitness. With the resources developed during this thesis, it may be possible to reassess this study using pedigree (**Chapter 2**) and/or genome resequencing resources (**Chapter 5**) in lieu of microsatellites. Further, these genomic and pedigree resources available for kakī could be used along with quantitative trait locus (QTL) mapping and genome-wide association scans (GWAS; Slate *et al.* 2009; Harrison *et al.* 2014; Santure & Garant 2018) to elucidate whether there are genomic regions associated with traits, such as reproductive success in kakī. It should be noted that questions regarding the genomic basis of reduced fertility have been notoriously difficult to assess in small populations with few genomic resources, as the population size is often too small for a high degree of statistical power using genome-wide association studies (Kardos *et al.* 2016b) and the phenotype of reduced fertility is complex and may be linked to multiple genomic regions (Kardos & Shafer 2019). We anticipate this research is still

worthwhile to explore given the resources available for kakī, if for no other reason than to reinforce the limitations of these approaches for genetically-depauperate species.

Beyond genome-wide approaches, we anticipate that targeted efforts will be more fruitful for understanding the genomic basis of traits in small populations. For example, exploring the Z-chromosome (i.e., large avian sex chromosome) as opposed to the entire genome may be an alternative approach for discovering the genomic basis of reduced fertility in birds, as many sex-specific characteristics are explained by variation found on this chromosome (Irwin 2018). For example, recent studies in zebra finch (*Taeniopygia guttata*) have shown that sperm morphology, and therefore motility, is largely influenced by a structural inversion on the Z-chromosome (Kim *et al.* 2017; Knief *et al.* 2017). For kakī and other critically endangered birds that experience reduced fertility (e.g., kākāpō, *Strigops habroptilus*, White *et al.* 2015; tara iti/fairy tern, *Sternula nereis davisae*, New Zealand Department of Conservation *unpublished data*), a targeted approach exploring the Z-chromosome may allow for more effective and efficient discovery of the genomic basis of reduced fertility.

While this research would provide a fantastic understanding of the genomic basis of fitness-related traits in small, genetically depauperate populations, relatively few researchers have addressed how this information can be adopted into best management practices for conservation breeding programmes. As described throughout this thesis, most programmes prioritise the retention of founder diversity, and in doing so, aim to ‘stop evolution’ by maintaining diversity and minimising any artificial selection of traits that are favourable in captivity (i.e., adaptation to captivity; Lacy 2009). If the genomic basis of maladaptive traits were elucidated, it is uncertain how selecting for or against such traits in a breeding programme would affect the retention of genome-wide diversity and recovery

success over several generations (Kardos & Shafer 2019). We anticipate for species facing threats that can cause imminent extinction (e.g., facial tumour disease in devils, Jones *et al.* 2007; chondrodystrophy in condors, Ralls *et al.* 2000; or chytrid fungus in many amphibians, Griffiths & Pavajeau 2008), selection for traits to prevent extinction will be weighed against the loss of genetic diversity resulting from selection. For example, a simulation study in field mice showed that selection for docility resulted in a significant reduction in diversity compared to a mean kinship strategy (Willoughby *et al.* 2015). Further, an analysis of California condor management strategies by Ralls *et al.* (2000) revealed that over half of the condor population would be removed from the breeding programme if all carriers of chondrodystrophy were selected against. This would be a substantial demographic loss to condors that would result in a severe reduction in gene diversity. To compromise, the breeding programme for condors has enacted a strategy to simultaneously breed birds with low kinship and avoid pairing potential heterozygous individuals, which maximises diversity while minimising the expression of chondrodystrophy (Ralls *et al.* 2000; Cynthia Steiner, *personal comm.*). Approaches such as these will likely need to be employed to evaluate best management options for intensively managed species with maladaptive traits. Because many whole genome resequencing data sets are becoming available not only for kakī and kākāriki karaka, but for other intensively managed species in our research group (i.e., tuturuatu/shore plover, *Thinornis novaeseelandiae*; tara iti/fairy tern, *Sternula nereis davisae*; kōwaro/Canterbury mudfish, *Neochanna burrowsius*), we envision an opportunity to use genome-wide simulations to explore whether selection against a maladaptive allele (or alleles in the case that the trait is polygenic) can be achieved while also maintaining genome-wide diversity in conservation breeding programmes. This research will suggest

whether functional genetic information can be incorporated into conservation breeding programmes, once the genomic basis of fitness traits have been elucidated.

## 6.6: Concluding Remarks

This thesis sits under the umbrella of large scale conservation genomics initiatives within the Conservation, Systematics, and Evolutionary Research Team (ConSERT, [www.uconsert.org/](http://www.uconsert.org/)) that use threatened species from Aotearoa to exemplify how to best maximise species resilience in the context of rapid global change. In addition to providing an exemplar for those transitioning from conservation genetics to conservation genomics, particularly for conservation breeding programmes, this thesis provides rich pedigree and genomic resources for both critically endangered kakī and kākāriki karaka, which will serve as the basis for continued research to enhance the recovery of both species. We are excited to see the approaches developed here for kakī and kākāriki karaka broadly applied beyond Aotearoa for threatened species recovery efforts worldwide.

## 6.7: References

- Attard CRM, Beheregaray LB, Möller LM (2018) Genotyping-by-sequencing for estimating relatedness in nonmodel organisms: Avoiding the trap of precise bias. *Molecular Ecology Resources*, **18**(3), 381–390.
- Ballou JD, Lacy RC (1995) Identifying genetically important individuals for management of genetic variation in pedigreed populations. *Population management for survival and recovery: analytical methods and strategies in small population conservation*. Columbia University Press: 76-111.
- Collier-Robinson L, Rayne A, Rupene M, Thoms C, Steeves T (2019) Embedding indigenous principles in genomic research of culturally significant species: a conservation genomics case study. New Zealand Ecological Society, *In Press*.
- De Villemereuil P, Rutschmann A, Lee KD, Ewen JG, Brekke P, Santure AW (2019) Little adaptive potential in a threatened passerine bird. *Current Biology*, **29**, 889-894.
- Dodds KG, McEwan JC, Brauning R, Anderson RM, van Stijn TC, Kristjánsson T, Clarke SM (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, **16**(1), 1047.

- Duchene DA, Bragg JG, Duchene S, Neaves LE, Potter S, Moritz C, Johnson RN, Ho SYW, Eldridge DB (2018) Analysis of phylogenomic tree space resolves relationships among marsupial families. *Systematic Biology*, **67(3)**, 400–412.
- Epstein B, Jones M, Hamede R, Hendricks S, McCallum H, Murchison EP, Schonfeld B, Wiench C, Hohenlohe P, Storfer A (2016) Rapid evolutionary response to transmissible cancer in Tasmanian devils. *Nature Communications*, **7**, 1–7.
- Escoda L, González-Esteban J, Gómez A, Castresana J (2016) Using relatedness networks to infer contemporary dispersal: Application to the endangered mammal *Galemys pyrenaicus*. *Molecular Ecology*, **26(13)**, 3343–3357.
- Frankham R (2008) Genetic adaptation to captivity in species conservation programs. *Molecular Ecology*, **17**, 325–333.
- Frankham R, Bradshaw CJA, Brook BW (2014) Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation*, **170**, 56–63.
- Galla SJ, Buckley TR, Elshire R, *et al.* (2016) Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances. *Molecular Ecology*, **25(21)**, 5267–5281.
- Galla SJ, Forsdick NJ, Brown L, *et al.* (2019). Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management. *Genes*, **10(1)**, 9.
- Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, **100(6)**, 659–674.
- Griffiths RA, Pavajeau L (2008) Captive breeding, reintroduction, and the conservation of amphibians. *Conservation Biology*, **22(4)**, 852–861.
- Grueber CE, Laws RJ, Nakagawa S, Jamieson IG (2010) Inbreeding depression accumulation across life-history stages of the endangered takahe. *Conservation Biology*, **24**, 1617–1625.
- Hagen EN, Hale ML, Maloney RF, Steeves TE (2011) Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the Black Stilt *Himantopus novaezelandiae*. *Ibis*, **153**, 556–561.
- Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014) Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, **7**, 1008–1025.



- He X, Johansson ML, Heath DD (2016). Role of genomics and transcriptomics in selection of reintroduction source populations. *Conservation Biology*, **30**(5), 1010–1018.
- Hellmann JK, Sovic MG, Gibbs HL, *et al.* (2016). Within-group relatedness is correlated with colony-level social structure and reproductive sharing in a social fish. *Molecular Ecology*, **25**(16), 4001–4013.
- Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, **93**, 47-64
- Hill WG, Weir BS (2012) Variation in actual relationship among descendants of inbred individuals. *Genetics Research*, **94**, 267–274.
- Hogg CJ, Wright B, Morris KM, Lee AV, Ivy JA, Grueber CE, Belov K (2018) Founder relationships and conservation management: empirical kinships reveal the effect on breeding programmes when founders are assumed to be unrelated. *Animal Conservation*, **22**(4), 348-361.
- Hubert J, Zerjal T, Hospital F (2018) Cancer- and behavior-related genes are targeted by selection in the Tasmanian devil (*Sacrophilus harrisii*). *PLoS one*, **13**(8), e0201838.
- Irwin DE (2018) Sex chromosomes and speciation in birds and other ZW systems. *Molecular Ecology*, **27**, 3831-3851.
- Ivy JA, Putnam AS, Navarro AY, Gurr J, Ryder OA (2016) Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, **107**(5), 403–412.
- Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, **115**(1), 63–72.
- Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW (2016a) Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, **9**, 1205-1218.
- Kardos M, Husby A, McFarlane SE, Qvarnstrom A, Ellegren H (2016b) Whole-genome resequencing of extreme phenotypes in collared flycatchers highlights the difficulty of detecting quantitative trait loci in natural populations. *Molecular Ecology Resources*, **16**, 727-741.
- Kardos M, Shafer ABA (2019) The peril of gene-targeted conservation. *Trends in Ecology and Evolution*, **33**(11), 827-839.
- Kim KW, Bennison C, Hemmings N, Brookes L, Hurley LL, Griffith SC, Burke T, Birkhead TR, Slate J (2017) A sex-linked supergene controls sperm morphology and swimming speed in a songbird. *Nature Ecology and Evolution*, **1**, 1168-1176.

- Knief U, Forstmeier W, Pei Y, *et al.* (2017) A sex-chromosome inversion causes strong overdominance for sperm traits that affect siring success. *Nature Ecology and Evolution*, **1**(8), 1177-1184.
- Lacy RC (2012) Extending pedigree analysis for uncertain parentage and diverse breeding systems. *Journal of Heredity*, **103**(2), 197–205.
- Lacy RC, Ballou JD, Pollak JP (2012) PMx: software package for demographic and genetic analysis and management of pedigreed populations. *Methods in Ecology and Evolution*, **3**, 433-437.
- Lazenby BT, Tobler MW, Brown WE, Hawkins CE, Hocking GJ, Hume F, Huxtable S, Iles P, Jones ME, Lawrence C, Thalmann S (2018) Density trends and demographic signals uncover the long-term impact of transmissible cancer in Tasmanian devils. *Journal of Applied Ecology*, **55**(3), 1368-1379.
- Lewin, HA, Robinson GE, Kress WJ, *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Science USA*, **115**(17), 4325-4333.
- Matasci N, Hung L, Yan Z, *et al.* (2014) Data access for the 1,000 Plants (1KP) project. *Gigascience*, **3**(1), 17.
- Peona V, Weissensteiner MH, Suh A (2018) How complete are “complete” genome assemblies?—An avian perspective. *Molecular Ecology Resources*, **18**, 1188-1195.
- Ralls K, Ballou JD, Rideout BA, Frankham R (2000) Genetic management of chondrodystrophy in California condors. *Animal Conservation*, **3**, 145-153.
- Robinson GE, Hackett KJ, Purcell-Miramontes M, *et al.* (2011) Creating a buzz about insect genomes. *Science*, **331**, 1386.
- Rhoads A, Au KF (2015) PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, **13**(5), 278-289.
- Romanov MN, Koriabine M, Nefedov M, de Jong PJ, Ryder OA (2006) Construction of a California condor BAC library and first-generation chicken-condor comparative physical map as an endangered species conservation genomics resource. *Genomics*, **88**, 711-718.
- Russello MA, Amato G (2004) Ex situ population management in the absence of pedigree information. *Molecular Ecology*, **13**, 2829-2840.
- Ryder O, Miller, W, Ralls K, Ballou JD, Steiner CC, Mitelberg A, Romanov MN, Chemnick LG, Mace, M, Schuster S (2016) Whole genome sequencing of California condors is now utilized for guiding genetic management. *International Plant and Animal Genome XXIV Conference*, 1.

- Santure AW, Garant D (2018) Wild GWAS—association mapping in natural populations. *Molecular Ecology Resources*, **18**, 729-738.
- Slate J, Gratten J, Beraldi D, Stapley J, Hale M, Pemberton JM (2009) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, **136**, 97-107.
- Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, **16**, 33–44.
- van Berkum NL, Lieberman-Aiden E, Williams L, *et al.* (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments*, **39**, e1869.
- Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, **2**, e41.
- Wang J (2016) Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, **107**, 4–13.
- White KL, Eason DK, Jamieson IG, Robertson BC (2015) Evidence of inbreeding depression in the critically endangered parrot, the kakapo. *Animal Conservation*, **18**, 341-347.
- Willoughby JA, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, Dewoody A (2015) The impacts of inbreeding, drift and selection on genetic diversity in conservation breeding populations. *Molecular Ecology*, **24**, 98-110.
- Wright B, Willet CE, Hamede R, Jones M, Belov K, Wade CM (2017) Variants in the host genome may inhibit tumour growth in devil facial tumours: evidence from genome-wide association. *Scientific Reports*, **7**, 423-429.
- Zhang G, Li B, Li C, Gilbert MTP, Jarvis ED, Wang J, The Avian Genome Consortium (2014) Comparative genomic data of the Avian Phylogenomics Project. *GigaScience*, **3**(1), 26.

## Appendix A: List of Conservation Genomics Reviews

Authors	Year	Title	Journal, Issue, Pages
Allendorf FW, Hohenlohe PA, Luikart G	2010	Genomics and the future of conservation genetics.	Nature Reviews Genetics, 11, 697– 709.
Allendorf FW	2017	Genetics and the conservation of natural populations: allozymes to genomes.	Molecular Ecology, 26, 420-430.
Amato G, DeSalle R, Ryder OA, Rosenbaum HC	2009	Conservation genetics in the age of genomics.	Columbia University Press.
Andrews KR, Luikart G	2014	Recent novel approaches for population genomics data analysis.	Molecular Ecology, 23(7), 1661-1667.
Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA	2016	Harnessing the power of RADseq for ecological and evolutionary genomics.	Nature Reviews Genetics, 17, 81–92.
Angeloni F, Wagermaker N, Vergeer P, Ouborg J	2011	Genomic toolboxes for conservation biologists.	Evolutionary Applications, 5, 130–143.
Avise JC	2008	The history, purview, and future of conservation genetics. In: Conservation Biology: Evolution in Action (eds Carroll SP, Fox CW).	Oxford University Press, Oxford. pp. 5–15.
Avise JC	2010	Perspective: conservation genetics enters the genomics era.	Conservation Genetics, 11, 665–669.
Benestan LM, Ferchaud L, Hohenlohe PA, <i>et al.</i>	2016	Conservation genomics of natural and managed populations: building a conceptual and practical framework.	Molecular Ecology, 25(13), 2967-2977.
Bonin A	2008	Population genomics: a new generation of genome scans to bridge the gap with functional genomics.	Molecular ecology, 17(16), 3583-3584.
Eklblom R, Galindo J	2011	Applications of next generation sequencing in molecular ecology of non-model organisms.	Heredity, 107, 1–15.
Ellegren H	2014	Genome sequencing and population genomics in non-model organisms.	Trends in Ecology and Evolution, 29, 51–63.
Frankham R	2010a	Where are we in conservation genetics and where do we need to go?	Conservation Genetics, 11, 661–663.
Frankham R	2010b	Challenges and opportunities of genetic approaches to biological conservation.	Biological Conservation, 143, 1919–1927.
Fraser DJ, Bernatchez L	2001	Adaptive evolutionary conservation: towards a unified concept for defining conservation units.	Molecular Ecology, 10(12), 2741-2752.
Fuentes-Pardo AP, Ruzzante DE	2017	Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations.	Molecular Ecology, 26, 5369-5406.
Funk WC, McKay JK, Hohenlohe PA, Allendorf FW	2012	Harnessing genomics for delineating conservation units.	Trends in Ecology and Evolution, 27, 489–496.
Galla SJ, Buckley TR, Elshire R, <i>et al.</i>	2016	uilding strong relationships between conservation genetics and primary industry	Molecular Ecology, 25, 5267-5281.

		leads to mutually beneficial genomic advances.	
Garner BA, Hand BK, Amish SJ, <i>et al.</i>	2016	Genomics in conservation: case studies and bridging the gap between data and application.	Trends in Ecology and Evolution, 31, 81–83.
Garvin MR, Saitoh K, Gharrett, AJ	2010	Application of single nucleotide polymorphisms to non-model species: a technical review.	Molecular Ecology Resources, 10(6), 915-934.
Gompert Z	2012	Population genomics as a new tool for wildlife management.	Molecular Ecology, 21(7), 1542-1544.
González-Martínez SC, Krutovsky KV, Neale DB	2006	Forest-tree population genomics and adaptive evolution.	New Phytologist, 170(2), 227-238.
Grueber CE	2016	Comparative genomics for biodiversity conservation.	Computational and Structural Biotechnology, 13, 370– 375.
Hand BK, Lowe WH, Kovach RP, Muhlfeld CC, Luikart G	2015	Landscape community genomics: understanding eco-evolutionary processes in complex environments.	Trends in Ecology & Evolution, 30(3), 161-168.
Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P	2014	Using genomics to characterize evolutionary potential for conservation of wild populations.	Evolutionary Applications, 7, 1008–1025.
He X, Johansson ML, Heath DD	2016	Role of genomics and transcriptomics in selection of reintroduction source populations.	Conservation Biology, 30, 1010–1018.
Hemmer-Hansen J, Therkildsen NO, Pujolar JM	2014	Population genomics of marine fishes: next-generation prospects and challenges.	The Biological Bulletin, 227(2), 117-132.
Hendricks S, Anderson EC, Antao T, <i>et al.</i>	2018	Recent advances in conservation and population genomics data analysis.	Evolutionary Applications, 11(8), 1197-1211.
Hidalgo E, González-Martínez SC, Lexer C, Heinze B	2010	Conservation genomics. In Genetics and Genomics of <i>Populus</i> .	Springer, New York, NY. p. 349-368.
Hoffmann A, Griffin P, Dillon S, <i>et al.</i>	2015	A framework for incorporating evolutionary genomics into biodiversity conservation and management.	Climate Change Responses, 2(1), 1.
Holliday JA, Aitken SN, Cooke JE, <i>et al.</i>	2017	Advances in ecological genomics in forest trees and applications to genetic resources conservation and breeding.	Molecular Ecology, 26(3), 706-717.
Johnson JA, Altwegg R, Evans DM, <i>et al.</i>	2016	Is there a future for genome-editing technologies in conservation?	Animal Conservation, 19, 97–101.
Johnson WE, Koepfli K	2014	The role of genomics in conservation and reproductive sciences.	Springer, New York, NY, 71-96.
Jones MR, Good JM	2016	Targeted capture in evolutionary and ecological genomics.	Molecular Ecology, 25, 185–202.
Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW	2016	Genomics advances the study of inbreeding depression in the wild.	Evolutionary Applications, 9(10), 1205-1218.
Khan S, Nabi G, Ullah MW, Yousaf M, Manan S, Siddique R, Hou H	2016	Overview on the role of advance genomics in conservation biology of endangered species.	International Journal of Genomics, 2016.

Kohn MH, Murphy WJ, Ostrander EA, Wayne RK	2006	Genomics and conservation genetics.	Trends in Ecology and Evolution, 21, 629–637.
Kraus RH, Wink M	2015	Avian genomics: fledging into the wild!	Journal of Ornithology, 156(4), 851-865.
Lozier JD, Zayed A	2017	Bee conservation in the age of genomics.	Conservation Genetics, 18(3), 713-729.
Luikart G, England PR, Tallmon D, Jordan S, Taberlet P	2003	The power and promise of population genomics: from genotyping to genome typing.	Nature Reviews Genetics, 4, 981–994.
McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT	2013	Applications of next-generation sequencing to phylogeography and phylogenetics.	Molecular Phylogenetics and Evolution, 66, 526–538.
McMahon BJ, Teeling EC, Ho€glund J	2014	How and why should we implement genomics into conservation?	Evolutionary Applications, 7, 999–1007.
Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA	2013	Genotyping-by-sequencing in ecological and conservation genomics.	Molecular Ecology, 22, 2841–2847.
Neale DB, Kremer A	2011	Forest tree genomics: growing resources and applications.	Nature Reviews Genetics, 12, 111–112.
Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D	2009	Population genomics of marine fishes: identifying adaptive variation in space and time.	Molecular Ecology, 18(15), 3128-3150.
Ouborg NJ	2010	Integrating population genics and conservation biology in the era of genomics.	Biology Letters, 6, 3-6.
Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW	2010	Conservation genetics in transition to conservation genomics.	Trends in Genetics, 26, 177–187.
Ouborg NJ, Angeloni F, Vergeer P	2010	An essay on the necessity and feasibility of conservation genomics.	Conservation Genetics, 11(2), 643-653.
Oyler-McCance SJ, Oh KP, Langin KM, Aldridge CL	2016	A field ornithologist's guide to genomics: practical considerations for ecology and conservation.	The Auk, 133(4), 626-648.
Perry GH	2014	The promise and practicality of population genomics research with endangered species.	International Journal of Primatology, 35(1), 55-70.
Piccolo JJ	2016	Conservation genomics: coming to a salmonid near you.	Journal of Fish Biology, 89(6), 2735-2740.
Price AL, Zaitlen NA, Reich D, Patterson N	2010	New approaches to population stratification in genome-wide association studies.	Nature Review Genetics, 11, 459–463.
Primmer CR	2009	From conservation genetics to conservation genomics.	Annals of the New York Academy of Science, 1162, 357–368.
Procaccini G, Olsen JL, Reusch TB	2007	Contribution of genetics and genomics to seagrass biology and conservation.	Journal of Experimental Marine Biology and Ecology, 350, 234-259.
Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM	2013	Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics.	Molecular Ecology, 22(11), 2953-2970.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R	2015	A practical guide to environmental association analysis in landscape genomics.	Molecular Ecology, 24(17), 4348-4370.
Romanov MN, Tuttle EM, Houck ML, <i>et al.</i>	2009	The value of avian genomics to the conservation of wildlife.	BMC Genomics, 10(2), 2009.
Russello MA, Waterhouse MD, Etter PD, Johnson EA	2015	From promise to practice: pairing non-invasive sampling with genomics in conservation.	PeerJ, 3, e1106.
Ryder OA	2005	Conservation genomics: applying whole genome studies to conservation efforts.	Cytogenetic and Genome Research, 108, 6-15.
Savolainen O, Lascoux M, Meril�a J	2013	Ecological genomics of local adaptation.	Nature Reviews Genetics, 14, 807–820.
Schielzeth H, Husby A	2014	Challenges and prospects in genome-wide quantitative trait loci mapping of standing genetic variation in natural populations.	Annals of the New York Academy of Sciences, 1320, 35–57.
Schwartz MK, McKelvey KS, Cushman SA, Luikart G	2010	Landscape genomics: a brief perspective. In Spatial complexity, informatics, and wildlife conservation.	Springer, Tokyo, 165-174.
Shafer ABA, Wolf JBW, Alves PC, <i>et al.</i>	2015	Genomics and the challenging translation into conservation practice.	Trends in Ecology and Evolution, 30, 78–87.
Shafer ABA, Wolf JBW, Alves PC, <i>et al.</i>	2016	Reply to Garner et al.	Trends in Ecology and Evolution, 31, 83–84.
Shaffer HB, Gidi� M, McCartney-Melstad E, Neal KM, Oyamaguchi HM, Tellez M, Toffelmier EM	2015	Conservation genetics and genomics of amphibians and reptiles.	Annual Review Animal Bioscience, 3, 113-138.
Stapley J, Reger J, Feulner PGD, <i>et al.</i>	2010	Adaptation genomics: the next generation.	Trends in Ecology and Evolution, 25, 705–712.
Steiner CC, Putnam AS, Hoeck PEA, Ryder OA	2013	Conservation genomics of threatened animal species.	Annual Review of Animal Bioscience, 1, 261–281.
Taylor HR, Gemmell NJ	2016	Emerging technologies to conserve biodiversity: further opportunities via genomics. Response to Pimm et al.	Trends in Ecology & Evolution, 31(3), 171-172.
Tigano A, Friesen VL	2016	Genomics of local adaptation with gene flow.	Molecular Ecology, 25, 2144–2164.
Todd EV, Black MA, Gemmell NJ	2016	The power and promise of RNA-seq in ecology and evolution.	Molecular Ecology, 25, 1224–1241.
Travers SE, Smith MD, Bai J, <i>et al.</i>	2007	Ecological genomics: making the leap from model systems in the lab to native populations in the field.	Frontiers in Ecology and the Environment, 5(1), 19-24.
Verhoeven KJF, vonHoldt BM, Sork VL	2016	Epigenetics in ecology and evolution: what we know and what we need to know.	Molecular Ecology, 25, 1631–1638.
vonHoldt BM, Brzeski KE, Wilcove DS, Rutledge LY	2018	Redefining the role of admixture and genomics in species conservation.	Conservation Letters, 11(2), e12371.

# Appendix B: Chapter 3, Published Manuscript

## MOLECULAR ECOLOGY

Molecular Ecology (2016) 25, 5267–5281

### NEWS AND VIEWS

#### OPINION

#### Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances

STEPHANIE J. GALLA,\*  
THOMAS R. BUCKLEY,†‡ ROB ELSHIRE,§  
MARIE L. HALE,\* MICHAEL KNAPP,¶ JOHN  
MCCALLUM,\*\* ROGER MORAGA,††  
ANNA W. SANTURE,‡ PHILLIP WILCOX‡‡  
and TAMMY E. STEEVES\*

\*School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand; †Landcare Research, Private Bag 92170, Auckland Mail Centre, Auckland 1142, New Zealand; ‡School of Biological Sciences, University of Auckland, Auckland 1010, New Zealand; §The Elshire Group, Ltd., 52 Victoria Avenue, Palmerston North 4410, New Zealand; ¶Department of Anatomy, University of Otago, P.O. Box 913, Dunedin 9054, New Zealand; \*\*Breeding and Genomics, New Zealand Institute for Plant and Food Research, Private Bag 4704, Christchurch 8140, New Zealand; ††AgResearch, Ruakura Research Centre, Bisley Road, Private Bag 3115, Hamilton 3240, New Zealand; ‡‡Department of Mathematics and Statistics, University of Otago, P.O. Box 56, 710 Cumberland Street, Dunedin 9054, New Zealand

Several reviews in the past decade have heralded the benefits of embracing high-throughput sequencing technologies to inform conservation policy and the management of threatened species, but few have offered practical advice on how to expedite the transition from conservation genetics to conservation genomics. Here, we argue that an effective and efficient way to navigate this transition is to capitalize on emerging synergies between conservation genetics and primary industry (e.g., agriculture, fisheries, forestry and horticulture). Here, we demonstrate how building strong relationships between conservation geneticists and primary industry scientists is leading to mutually-beneficial outcomes for both disciplines. Based on our collective experience as collaborative New Zealand-based scientists, we also provide insight for forging these cross-sector relationships.

**Keywords:** applied research, conservation genomics, high-throughput sequencing, interdisciplinary research, next-generation sequencing

Correspondence: Stephanie J. Galla, Fax: +64 3 364 2590; E-mail: stephanie.galla@pg.canterbury.ac.nz

Received 22 April 2016; revision received 23 August 2016; accepted 24 August 2016

One does not need to read beyond the pages of *Molecular Ecology* to see how emerging technologies are revolutionizing the way we conduct research in ecology and evolutionary biology (i.e. EEB) and conservation biology. This is exemplified by rapid advances in genomics, where in the span of two decades the field of molecular ecology has grown from using Sanger technologies to sequence single target loci to using high-throughput sequencing (HTS) technologies to affordably sequence entire draft genomes (Narum *et al.* 2013; Payseur & Rieseberg 2016; Tigano & Friesen 2016). When new technologies become available, there is a tendency for reviews to be published heralding their potential to address new and exciting questions. Beyond the value of these reviews, an even more important conversation needs to take place in the peer-reviewed literature: How do we efficiently incorporate new technologies into our research repertoire to make accelerated gains in applied and fundamental science?

The field of conservation genetics is currently in transition given rapid advancements in HTS technologies. Many reviews have highlighted the promise of embracing HTS technologies in conservation (Luikart *et al.* 2003; Kohn *et al.* 2006; Primmer 2009; Allendorf *et al.* 2010; Avise 2010; Frankham 2010a; Ouborg *et al.* 2010; Angeloni *et al.* 2011; Ekblom & Galindo 2011; Funk *et al.* 2012; McCormack *et al.* 2013; Narum *et al.* 2013; Steiner *et al.* 2013; Ellegren 2014; McMahon *et al.* 2014; Shafer *et al.* 2015; Andrews *et al.* 2016; Benestan *et al.* 2016; Grueber 2016). However, as recently discussed by Shafer *et al.* (2015, 2016) and Garner *et al.* (2016), there are a limited (albeit increasing) number of published empirical studies that apply HTS data to conservation. We are aware of empirical genomic studies in EEB that are applicable to questions in conservation (e.g. Defaveri *et al.* 2013; Hoffman *et al.* 2014; Knief *et al.* 2015; Bérénos *et al.* 2016; Hess *et al.* 2016; Prince *et al.* 2016) and there are many EEB researchers applying their genomics expertise to improve conservation outcomes for threatened species, including two of our co-authors (MK, AWS). In addition to the EEB sphere, there are conservation geneticists (e.g. our co-authors SJG, TRB, MLH, TES) who are successfully venturing into conservation genomics through collaborations with colleagues in another applied discipline well versed in genomics: primary industry (a collective term referring to scientists in agriculture, fisheries, forestry and horticulture; such as our co-authors RE, JM, RM, PW). Through building these cross-sector relationships, it has become clear that there is immense potential for conservation geneticists and primary industry scientists to collaborate on applied research that addresses aligned questions using similar genomic approaches. In this opinion piece, we use our experience as a collaborative group of New

© 2016 The Authors. *Molecular Ecology* Published by John Wiley & Sons Ltd.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.



Zealand-based scientists to argue that building strong relationships between conservation genetics and primary industry can lead to improved genomic outcomes for both disciplines and offer advice on how to best build meaningful cross-sector relationships.

### Conservation genetics and genomics

Before discussing mutually beneficial genomic synergies between conservation genetics and primary industry, we feel it is important to first address what conservation genetics is, what can be gained using a genomic approach and what obstacles may impede geneticists from adopting genomic technologies. Conservation genetics is a subdiscipline of conservation biology (Soulé 1985) which uses genetic data to inform the management of threatened species in collaboration with conservation practitioners (Frankham 1995, 2010b; Avise 2008; Haig *et al.* 2016). While there is overlap between the fields of conservation genetics and EEB, we distinguish conservation genetics as an applied subdiscipline with direct implications for the management of threatened species. Many threatened taxa have experienced significant population declines (i.e. demographic bottlenecks, see Keller *et al.* 1994), leading to small populations that are susceptible to genetic factors (i.e. loss of genetic diversity, inbreeding and inbreeding depression) associated with extinction risk (Frankham 1995). Conservation geneticists have traditionally used few targeted neutral genetic markers including mitochondrial sequences, microsatellites and amplified fragment length polymorphisms (AFLPs) to measure inbreeding, relatedness and genetic diversity within threatened populations, estimate population genetic structure and gene flow among threatened populations, delineate species boundaries in threatened taxa and detect hybridization and introgression between threatened and nonthreatened species (Allendorf *et al.* 2010; Ouborg *et al.* 2010).

Advancements in HTS technologies are enabling the development of genomic resources for threatened species including the *de novo* assembly and annotation of high-quality reference genomes (e.g. Li *et al.* 2014; Zhang *et al.* 2014) and characterization of a large number of genome-wide markers such as single nucleotide polymorphisms (SNPs) (e.g. Benestan *et al.* 2015; Kraus *et al.* 2015; Lemay & Russello 2015). For conservation geneticists who have traditionally used small panels of neutral genetic markers to estimate population genetic parameters above and below the species level, HTS technologies are appealing as they enable an affordable means to discover and genotype a large quantity of genome-wide SNPs (Avise 2010; McCormack *et al.* 2013; Shafer *et al.* 2015) and these large SNP data sets are more representative of genome-wide variation and can result in higher resolution estimates of population genetic parameters (Väli *et al.* 2008; Ljungqvist *et al.* 2010; Santure *et al.* 2010; Taylor *et al.* 2015). In the field of conservation genetics and EEB, a small but rapidly growing number of empirical studies have demonstrated the utility of genomic markers in estimating population genetic structure and gene flow (Bowden *et al.* 2012; Dierickx *et al.* 2015; Lew *et al.* 2015;

Oyler-McCance *et al.* 2015), estimating relatedness (Béréanos *et al.* 2016), measuring genome-wide diversity (Robinson *et al.* 2016) and detecting hybridization and introgression (Hohenlohe *et al.* 2011). We anticipate even more conservation geneticists will begin to embrace HTS technologies as empirical evidence demonstrating the superiority of using genomic markers to inform conservation decisions grows and the costs of doing so diminishes (Box 1).

The paradigm underlying many conservation genetic studies is that a genetically diverse population as measured by neutral genetic markers is also likely to be functionally diverse (Bataillon *et al.* 1996) and therefore better able to adapt to environmental change (Frankham 2005). While many have aspired to move past this paradigm, it remains entrenched in most conservation genetic studies that use neutral markers (Caballero & García-Dorado 2013; Vilas *et al.* 2015). As a result of the lack of empirical data on functional genetic diversity in species of conservation interest, beyond studies that include immunocompetence genes like those in the major histocompatibility complex and Toll-like receptors (reviewed in Grueber 2016), it has been difficult to assess the validity of this conservation genetic paradigm. Further, even if supported by empirical data, neutral genetic data might not be a suitable proxy for functional genetic data for threatened species. For example, the translocation of individuals from a large genetically diverse population to supplement a small genetically depauperate population might introduce new genetic diversity (Weeks *et al.* 2011; IUCN/SSC 2013), but it might also inadvertently lead to outbreeding depression if source and recipient populations are each locally adapted (Edmands 2007; Frankham *et al.* 2011; but see Frankham 2015; Whiteley *et al.* 2015; He *et al.* 2016).

There is exceptional interest in using a conservation genomics approach to detect regions of the genome that underlie phenotypic variation linked to fitness in threatened populations (i.e. adaptive variation; Luikart *et al.* 2003; Kohn *et al.* 2006; Ouborg *et al.* 2010; Angeloni *et al.* 2011; Harrison *et al.* 2014; Shafer *et al.* 2015). There are several methods available to study adaptive variation, including gene-mapping approaches (i.e. genome-wide association studies or GWAS, and quantitative trait loci mapping or QTL; Slate *et al.* 2010; Stapley *et al.* 2010), outlier locus analysis (Luikart *et al.* 2003; Haas & Payseur 2016) and selective sweep mapping (Pardo-Diaz *et al.* 2015). However, determining the genetic basis of phenotypic traits, especially those linked to fitness, is complex, owing to the fact that most fitness-related traits are likely to be controlled by multiple loci (Savolainen *et al.* 2013) and many are likely to be under at least some environmental influence (Falconer & Mackay 1996; Lynch & Walsh 1998). In addition, the success of these approaches is often contingent on large sample sizes (e.g. Ball 2005) which will be challenging to generate for most species of conservation concern.

While there are challenges associated with the detection of adaptive variation in threatened populations (reviewed in Shafer *et al.* 2015), there is potential to answer new questions previously not tractable by employing small sets of targeted genetic markers. In particular, an understanding

**Box 1.** The costs of using a conservation genomic approach. Perspectives are those of Tammy Steeves.

Since I arrived in New Zealand from Canada in 2004, I have had the privilege of developing conservation genetic management recommendations in collaboration with several Department of Conservation recovery or specialist groups to assist the recovery of endemic taonga (treasured) bird species. To date, these recommendations have been predominantly based on genetic markers, namely mitochondrial sequences or microsatellite genotypes (e.g. Steeves *et al.* 2010; Hagen *et al.* 2011; Overbeek *et al.* 2016). In collaboration with primary industry colleagues in the MapNet community (see Boxes 2 & 3), I recently assessed the direct and indirect costs associated with shifting from a conservation genetic to a conservation genomic approach and decided to develop genomic markers (SNPs) for the endangered tuturuatu/shore plover (*Thinornis novaeseelandiae*; Fig. A) and the critically endangered kaki/black stilt (*Himantopus novaeseelandiae*; Fig. B). [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)].



**Fig. A** Tuturuatu/Shore plover

*Tuturuatu/Shore plover*—I was recently invited to be an expert advisor to the Shore Plover Specialist Group. The Specialist Group was interested in sampling captive and wild birds to estimate the extent of population genetic structure and compare levels of genetic diversity, between captive and wild tuturuatu. To achieve this, I knew the cost to develop, screen and genotype ~20 polymorphic species-specific microsatellites for 94 individuals (~10K NZD) would be more than using a reduced-representation approach to simultaneously discover and genotype >20 000 SNPs for the same number of individuals (Elshire *et al.* 2011; ~8.5K NZD). I also knew it would be possible to expedite the characterization of SNPs if I was able to use a reference-guided approach. As a member of the Avian Genome Consortium, I was aware bird genomes are small, compact and highly conserved (Zhang *et al.* 2014), and that one of the newly available high-quality bird genomes (killdeer, *Charadrius vociferus*) would likely be an appropriate proxy-reference genome for SNP discovery and genotyping in tuturuatu because both

species are members of the Family Charadriidae (Card *et al.* 2014). Thus, the main driver of my decision to embrace a conservation genomic approach for tuturuatu was to ensure that I could develop a comprehensive postgraduate research project that could deliver pertinent results to the Shore Plover Specialist Group in a timely fashion.



**Fig. B** Kaki/Black stilt

*Kaki/Black stilt* – As a member of the Kaki Recovery Group, I have used species-specific genetic markers to inform the conservation genetic management of captive and wild kaki populations for many years. For example, I routinely use genetic-based measures of relatedness based on microsatellites to inform captive pairing decisions (as per Hagen *et al.* 2011). However, emerging evidence indicates genetic-based measures are relatively poor indicators of genome-wide diversity, particularly in genetically impoverished species like kaki, and a better indication of genome-wide diversity should be obtained from genomic-based measures of relatedness based on genome-wide SNPs (Taylor *et al.* 2015; Willoughby *et al.* 2015). Thus, the main driver of my decision to generate SNPs for kaki was to establish the Kaki Recovery Programme as an exemplar of 'best practice' conservation genomic management.

of the genetic basis of fitness traits will allow more robust predictions of the evolutionary potential of threatened species (Ouborg *et al.* 2010; Harrison *et al.* 2014), including a better understanding of genetic trade-offs between traits that might constrain adaptation (Slate *et al.* 2010). Further, identifying loci underlying local adaptation is likely to help identify candidate populations for conservation translocations (Seddon 2010; He *et al.* 2016). Finally, identification of

genes responsible for detrimental traits associated with inbreeding depression will have immediate impact on the management of threatened species, especially where matings between individuals are managed (e.g. captive populations; Angeloni *et al.* 2011; Harrison *et al.* 2014; Shafer *et al.* 2015).

Despite having been available for over a decade (Margulies *et al.* 2005), a limited number of publications

have applied HTS technologies to conservation (Shafer *et al.* 2015, 2016; but see Garner *et al.* 2016), with the term 'conservation genomics gap' first being used in 2015 to describe the paucity of conservation geneticists using HTS technologies to inform conservation management (Shafer *et al.* 2015). While there are a growing number of examples that show how genomic data are being used to inform conservation decisions (Garner *et al.* 2016; but see Shafer *et al.* 2016; see Fig. S1, Supporting information) and many conservation geneticists who are currently producing HTS data sets, there has been a substantial time lag between when these techniques have become available and uptake by the conservation research community, especially in comparison with other applied genetic disciplines like primary industry (e.g. agriculture, fisheries, forestry and horticulture; see Fig. 1). In addition, much of the uptake in conservation biology has been restricted to threatened wild fish stocks (Garner *et al.* 2016; Shafer *et al.* 2016). Of the 51 articles in Fig. 1 classified as 'conservation genomics', 30% pertained to the management of declining, overfished or threatened commercially fished species (e.g. Atlantic salmon, *Salmo salar*; orange roughy, *Hoplostethus atlanticus*; delta smelt, *Hypomesus transpacificus*), which provides an excellent example of how conservation genomic research can also be relevant to other applied scientific disciplines including primary industry (e.g., these articles were classified as both 'conservation genomics' and 'primary industry' in Fig. 1).

Shafer *et al.* (2015) predominantly attribute the conservation genomics gap to a persistent disconnect between academia and real-world conservation issues. We agree strong relationships between academics and conservation practitioners are crucial, but argue the conservation genomics gap as defined by Shafer *et al.* (2015) is more akin to a

'research-implementation gap' (Knight *et al.* 2008; Hogg *et al.* 2016). Indeed, if strong relationships between academics and conservation practitioners are absent, the likelihood that *any* research will be translated into conservation action is exceptionally low (Haig *et al.* 2016). Here, we predominantly attribute the apparent shortage of conservation geneticists using HTS technologies (i.e. the conservation genomics gap *sensu stricto*) to several interconnected challenges associated with the generation, analysis and interpretation of genomic data.

Prior to identifying these interconnected challenges, we recognize some questions in conservation are still being readily addressed with genetic data (e.g. Dowling *et al.* 2015; Li *et al.* 2015a; Pacioni *et al.* 2015; Trask *et al.* 2015; Cubrinovska *et al.* 2016; Hammerly *et al.* 2016; Overbeek *et al.* 2016). We anticipate studies such as these to persist, at least in the short term, because existing panels of genetic markers remain a sufficient low-cost option in some situations (Angeloni *et al.* 2011; McCormack *et al.* 2013; McMahon *et al.* 2014). Although we acknowledge that direct cost can be a factor contributing to the conservation genomics gap, we do not think it underpins it, especially when reduced-representation approaches (e.g. restriction site-associated DNA sequencing, genotyping-by-sequencing, exome capture and RAD capture; Baird *et al.* 2008; Elshire *et al.* 2011; Jones & Good 2016; Ali *et al.* 2016) make it possible to characterize tens of thousands of SNPs in hundreds of individuals for nonmodel species at a lower cost than developing and screening relatively few novel microsatellite markers (Narum *et al.* 2013; Andrews *et al.* 2016; Box 1). Beyond direct cost, the shortage of high-quality reference genomes is an often cited impediment to SNP discovery and genotyping for nonmodel species (e.g. Allendorf *et al.* 2010; Ouborg *et al.* 2010; Shafer *et al.* 2015), particularly when approximate SNP location is of interest (e.g. Kardos *et al.* 2015). However, an ever increasing number of high-quality and high-coverage genomes are becoming available (Ellegren 2014). It has also become apparent that low-coverage draft genomes (sometimes referred to as 'landing-pad' or 'skim' genomes), or even high-quality and high-coverage genomes of closely related taxa, can enable reference-guided mapping assembly and SNP characterization in some taxa (Card *et al.* 2014; Wang *et al.* 2014). The lack of bioinformatic expertise and pipelines required to analyse large population genomic data sets has also been frequently cited as a challenge that precludes the use of HTS technologies in conservation (e.g. McCormack *et al.* 2013; Shafer *et al.* 2015). Steep analytical learning curves are generally associated with new technologies, particularly for rapidly advancing fields like genomics where bioinformatic expertise is needed to analyse large genomic data sets. However, the analysis of large population genomic data sets is no longer exceptional. For example, in regard to SNP discovery and genotyping alone, several comprehensive bioinformatic pipelines are readily available (e.g. Glaubitz *et al.* 2014; Puritz *et al.* 2014; Herten *et al.* 2015; Sovic *et al.* 2015; Melo *et al.* 2016).

Depending on the conservation genetics project at hand, one or a combination of the challenges listed above might

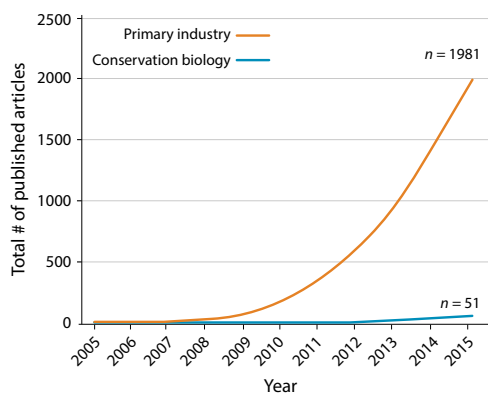


Fig. 1 Number of publications using high-throughput sequencing technologies to generate genomic data in conservation (blue line) and primary industry (red line) from 2005 to 2015. Values for this graph were derived from an ISI Web of Science literature search, using inclusive terminology (see Data S1, Supporting information for details). Curve lines have been smoothed for ease of interpretation. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)].

impede conservation geneticists from transitioning to HTS technologies. Given the recent developments in HTS technologies and the potential it has for benefitting conservation outcomes, we suggest it is time for researchers to start sharing practical advice on how to expedite the transition from conservation genetics to conservation genomics. Here, we argue that an effective and efficient way to navigate the conservation genomics gap is to capitalize on emerging synergies between conservation genetics and primary industry, and demonstrate how building strong relationships between these two disciplines is leading to mutually beneficial genomic outcomes.

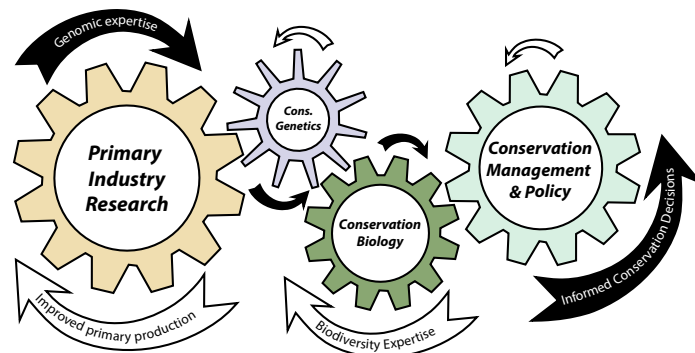
### Strong relationships lead to mutually beneficial genomic advances

Conservation geneticists are skilled at building strong relationships in an interdisciplinary landscape to improve conservation outcomes (Haig *et al.* 2016; Hogg *et al.* 2016). However, by pushing the boundaries of the conservation 'silo', conservation geneticists will be better able to navigate the conservation genomics gap if they forge novel relationships with scientists that have shared genomic goals, albeit in a different discipline such as primary industry (Fig. 2). As a discipline, primary industry represents a diverse group of scientists from universities, private institutions and government organizations that apply scientific data to the benefit of primary production output (e.g. meat, fish, eggs, dairy, fruits, vegetables, fibres and timber). Some of the early draft genomes were published to improve commercial outcomes, including rice (*Oryza sativa*; Goff *et al.* 2002), red jungle fowl (*Gallus gallus*; Hillier *et al.* 2004), silkworm (*Bombyx mori*; Xia *et al.* 2004) and cattle (*Bos taurus*; Schibler *et al.* 2004). With these early reference genomes and the accumulation of massive SNP data sets coupled with phenotypic data, many primary industry scientists have years of expertise with the application of genomic data. Approximately 1981 HTS studies using genomic data have been published in primary industry from 2005 to 2015, which outnumbers those produced in conservation biology by more than an order of magnitude (Fig. 1).

Conservation has already benefitted from genomic resources provided by primary industry. For example,

genomic resources developed for cattle including the draft genome (Schibler *et al.* 2004) and the bovine SNP chip (Gunderson *et al.* 2005; Steemers *et al.* 2006; Matukumalli *et al.* 2009) have been used to estimate the extent of introgression from cattle to American bison (*Bison bison*; Halbert *et al.* 2005), measure genomic variation in American and European bison (*B. bonasus*; Pertoldi *et al.* 2009) and develop genomic resources for scimitar-horned and Arabian oryx (*Oryx dammah* and *O. leucoryx*, respectively; Ogden *et al.* 2012). Similarly, genomic resources developed for domestic sheep (*Ovis aries*) have been used to describe genome-wide diversity and assess genetic rescue for bighorn sheep (*Ovis canadensis*; Poissant *et al.* 2009; Miller *et al.* 2012). Of course, there are species of mutual interest to both conservation and primary industry, including species in the fishery and forestry sectors (e.g. Monterey pine, *Pinus radiata* D. Don; New Zealand tōtara, *Podocarpus* spp.; chinook salmon, *Oncorhynchus tshawytscha*; orange roughy, *Hoplostethus atlanticus*), and therefore, genomic resources produced by one discipline can be easily used by the other (Dillon *et al.* 2013; Larson *et al.* 2014; Marshall *et al.* 2015; da Silva *et al.* 2015). We anticipate conservation geneticists may opt to use closely related commercial or model species to inform adaptive variation studies in threatened species, given that genotyping approaches are contingent on large sample size (Ball 2005; see discussion above) and the small census size of threatened populations may be inadequate.

Collaborations between conservation geneticists and primary industry scientists are logical because researchers in these two disciplines are beginning to address similar questions in an applied genetic discipline (see Table 1). For example, primary industry scientists have been using neutral genome-wide SNPs to calculate inbreeding coefficients in domestic sheep (Li *et al.* 2011), reconstruct parentage assignments in cattle (Hayes 2011) and calculate diversity measures for genetic improvement in poultry (red jungle fowl, Muir *et al.* 2008; domestic turkey, *Meleagris gallopavo*, Aslam *et al.* 2012). Pipelines that have been used or developed to address these questions in commercial species are likely to be of interest to conservation geneticists, but are sometimes published in discipline-specific peer-reviewed journals such as the *Journal of Dairy Science* or *Plant Biotechnology Journal* (e.g. Allen *et al.* 2012; Li *et al.* 2015b).



**Fig. 2** Simplified schematic detailing how relationships between conservation genetics and primary industry are leading to mutually beneficial outcomes. In black arrows, genomic expertise from primary industry advances conservation genetics, which in turn informs conservation biology and conservation management and policy. In white arrows, biodiversity expertise informs primary industry research, which in turn improves primary production. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)].



**Table 1** Common genomic issues facing conservation genetics and parallel examples addressed by scientists in primary industry

Topic	Challenge for conservation genomics	Examples of corresponding research from primary industries
Polyploid genomes	Developing effective tools for genome-wide SNP discovery and genotyping for plants, invertebrates and some vertebrates with polyploid genomes	Genome-wide SNP studies on polyploids <sup>1</sup> including wheat <sup>2</sup> , cotton <sup>3</sup> , potato <sup>4</sup> and peanut <sup>5</sup>
Genetic basis of adaptive variants	Discovery of variants underpinning traits of relevance to conservation including adaptive variation	Trait mapping for economically important traits using GWAS and QTL mapping <sup>6,7</sup> in rice <sup>8</sup> , dairy cattle <sup>9</sup> , pig <sup>10</sup> and soy bean <sup>11</sup>
Gene copy number variation	Quantifying genome-wide copy number variation and estimating its contribution to phenotypic variation	Quantifying genome-wide copy number variation and estimating its contribution to economically important traits in apple <sup>12</sup> , pig <sup>13</sup> , wheat <sup>14</sup>
Inbreeding and relatedness	Measuring inbreeding ( $f$ ), detecting inbreeding depression and estimating relatedness ( $r$ ) for small populations to maintain evolutionary potential	Measuring inbreeding ( $f$ ), detecting inbreeding depression and estimating relatedness ( $r$ ) in sheep <sup>15</sup> , pigs <sup>16,17</sup> and salmon <sup>18</sup> to enhance traits for commercial selection

<sup>1</sup>Clevenger *et al.* (2015), <sup>2</sup>Allen *et al.* (2012), <sup>3</sup>Byers *et al.* (2012), <sup>4</sup>Uitdewilligen *et al.* (2013), <sup>5</sup>Bertioli *et al.* (2014), <sup>6</sup>Collard & Mackill (2008), <sup>7</sup>Hu *et al.* (2013), <sup>8</sup>Begum *et al.* (2015), <sup>9</sup>Li *et al.* (2015b), <sup>10</sup>Zhang *et al.* (2015), <sup>11</sup>Zhou *et al.* (2015), <sup>12</sup>Boocock *et al.* (2015), <sup>13</sup>Wang *et al.* (2015), <sup>14</sup>Wuerschum *et al.* (2015), <sup>15</sup>Li *et al.* (2011), <sup>16</sup>Herrero-Medrano *et al.* (2012), <sup>17</sup>Silió *et al.* (2016), <sup>18</sup>Dodds *et al.* (2015).

Similarly, there are some conservation genomic articles from nonacademic sources that are not represented in peer-reviewed literature (Garner *et al.* 2016). These examples highlight how relationships between conservation genetics and primary industry scientists can enable the dissemination of discipline-specific publications and will allow scientists from both disciplines to learn about recently developed pipelines.

Understanding the genetic basis of desired commercial traits is also a main focus in primary industry (Womack 2005; Tuberosa & Salvi 2006; Sellner *et al.* 2007; Collard & Mackill 2008; Neale & Kremer 2011; Sonah *et al.* 2011; Hu *et al.* 2013). Primary industry has benefitted from collaboration with researchers in human health to determine the genetic basis of phenotypic traits in complex pedigrees and structured populations using QTL mapping and GWAS (George *et al.* 2000; Aulchenko *et al.* 2007; Price *et al.* 2010). In turn, these gene-mapping approaches have been successfully applied to understanding the genetic basis of ecologically relevant traits in wild populations (Schielzeth & Husby 2014). While there are numerous research groups outside of primary industry exploring adaptive variation (e.g. Rietveld *et al.* 2013; Brachi *et al.* 2015; Chaves *et al.* 2016), we anticipate that conservation geneticists in particular will benefit from forging relationships with primary industry scientists given that both groups work in an applied discipline with species characterized by small effective population sizes. Additionally, there is potential for conservation geneticists to adopt a genomic selection approach (e.g. Hayes *et al.* 2009; Heffner *et al.* 2009) to generate breeding values to inform the selection of individuals for captive breeding. Lastly, we recognize that both conservation geneticists and primary industry researchers routinely work with species with complex genomes (Clevenger *et al.* 2015), and therefore researchers from these two disciplines have an opportunity to work together and think of creative bioinformatic solutions for species that present

bioinformatic challenges (Box 3). Given these commonalities, synergies between both conservation genetics and primary industry can lead to the development of improved HTS techniques and pipelines to address mutual problems in species of both conservation and commercial interest (Boxes 2 and 3; Table 1).

Relationships between conservation geneticists and primary industry scientists can result in improved commercial outcome for primary species as well. Conservation geneticists strive to preserve genetic diversity and the ecological and evolutionary processes that generate it (Groom *et al.* 2006; Haig *et al.* 2016). There is growing discussion among primary industry scientists regarding the need for commercial breeding programs to maximize genetic diversity and minimize inbreeding (Medugorac *et al.* 2009; Windig & Engelsma 2010; Joost *et al.* 2011; Lenstra *et al.* 2012; Pryce *et al.* 2012; Kristensen *et al.* 2015). Livestock and crops are often of a small effective population size (i.e.  $N_e < 100$ ) due to many generations of artificial selection for desired traits and are thus susceptible to loss of genome-wide variation via inbreeding and genetic drift (Windig & Engelsma 2010; Leroy *et al.* 2013; Kristensen *et al.* 2015; Jiménez-Mena *et al.* 2016; Shepherd *et al.* 2016). There is evidence for inbreeding depression in rare breeds, such as cashmere goats (*Capra aegagrus*; Dai *et al.* 2015), Iranian Guilan sheep (Eteqadi *et al.* 2015) and Iberian pigs (*Sus scrofa*; Saura *et al.* 2015). There is also an increasing awareness of the risks associated with deploying very few genotypes, particularly in the presence of novel crop pathogens (Kim *et al.* 2015) and an increasing concern among rare breeds regarding the loss of genetic variation associated with traits that might be useful in future markets (e.g. Catalanian donkey *Equus africanus*, Gutierrez *et al.* 2005; Famennoise poultry, Moula *et al.* 2009; black Slavonian pigs, Luković *et al.* 2012). Conservation geneticists have many years of expertise regarding the conservation genetic management strategies for threatened species (Frankham 2010a). As a result, conservation geneticists can

**Box 2.** Retrospective and prospective of genotyping by sequencing (GBS). Perspectives are those of Rob Elshire.

In 2007, I joined the Buckler Lab at Cornell University and the next-generation sequencing revolution simultaneously. My first task was to develop a new library preparation method for the nascent Illumina sequencing platform. The technology was not nearly as robust as it is today and the reads were very short (i.e. 32 bp in length). Our challenge was to sequence the nonrepetitive fraction of the maize (*Zea mays*) genome. To do that, we used a combination of digestion by restriction enzymes and gel-based size selection to exclude the repetitive fraction. The data generated formed the basis for the first Maize Hapmap paper (Gore *et al.* 2009). When that project neared completion, I was tasked with building a low-cost, high-throughput genotyping method as an extension of my previous work. The overall goal was to develop a genotyping system that would allow simultaneous marker discovery and genotyping and also address the issue of marker discovery bias. Other researcher groups at the time were developing similar methods, as there was a high demand for an affordable and reproducible method of genotyping and it was the next logical thing to try. One aim was to provide enough genetic markers at the right price point to enable plant breeding by genomic selection. To maximize the benefit of our work and encourage others to take what we did and create new methods appropriate for new questions, we made our work openly available. The resultant genotyping-by-sequencing (GBS) method was published in *PLoS One* in 2011 (Elshire *et al.* 2011).

We achieved our goal of developing a new genotyping method that was inexpensive, both in terms of cost per sample and cost per data point (i.e. fractions of a cent per marker). The low-cost and high-throughput nature of GBS allows plant breeders to genotype thousands of plants per cycle in genomic selection driven breeding programs (He *et al.* 2014). Primary industry programs in animal breeding have also taken up GBS. Unlike microsatellites or SNP chips, no previously generated genomic resources are necessary to deploy GBS. This allows researchers working in nonmodel species, such as orphan crops (i.e. crops of regional commercial importance, but not global), to take advantage of powerful genomic tools (Varshney *et al.* 2012). The situation for researchers in ecology and conservation biology is not dissimilar to that of those working with orphan crops. The budgets are small, resources meagre and the questions are of local importance with small (if any) obvious economic returns. It is no wonder that ecologists were among the earliest adopters of GBS.

During the development of the GBS, we tested it on species other than maize. Confident that it worked in a variety of kingdoms, we welcomed interested early adopters to the laboratory for assistance. Two of those early adopters worked in the ecology space. Dr. Thomas White worked with the invasive bank vole (*Myodes glareolus*, White *et al.* 2013) in Ireland which had small sample sizes and no reference genome. Dr. Nancy Chen studied the Florida scrub jay (*Aphelocoma coerulescens*) and developed a method using GBS data and Mendelian inheritance to improve SNP discovery (Chen *et al.* 2014). It became clear that we had developed a generally useful genomics research tool and it could be used by researchers across disciplines. We had already published the method in an open-access journal and provided analysis software under a free software licence. To allow researchers to more easily use this technology, we set up a GBS service at Cornell. By early 2016 the Cornell service had performed GBS analysis on over 1500 species.

After our initial GBS publication, a plethora of method modifications and additional software tools have emerged. The recently published epiGBS method (van Gurp *et al.* 2016) allows the interrogation of the methylome and does not require a reference genome, thereby extending the utility of the base method greatly. The GBSX toolkit (Herten *et al.* 2015) is a set of software designed to assist in the design of GBS based experiments. Many software packages have been developed to analyse GBS data (e.g. TASSEL-UNEAK, STACKS, GBS-SNP-CROPS, GLBPSS; Lu *et al.* 2013; Catchen *et al.* 2013; Hapke & Thiele 2016; Melo *et al.* 2016) that are appropriate for species without reference genomes. Extensions to the molecular method and new software tools make these types of genomics approaches more broadly accessible; however, barriers to using this technology still exist in many disciplines, including the cost of laboratory and informatics setup and reservations in transitioning to new analytical tools.

Marker technology adoption has a long tail distribution. In 2013, I gave a talk on GBS at the *Molecular Markers in Horticulture Symposium*. Perusing the poster session, I found that researchers were using every type of marker technology that I knew about: from isozymes to GBS. Why were some researchers using cutting edge technologies? Why were others using antiquated, expensive and low information content technologies? Researchers in conservation genomics are in a similar situation. Across disciplines, the biological sciences are encountering rapidly changing technologies and increasingly larger data sets. Industry service providers with expert knowledge and experience, like my small New Zealand-based company (Elshire Group, Ltd.) and many others, can help bridge the gap. By developing relationships spanning human health, primary industry and conservation, as well as actively participating in research communities like MapNet (Box 3), we can work together to expedite the adoption of genomic technologies applicable to the questions at hand, effectively, efficiently and with confidence.

**Box 3.** Building strong interdisciplinary relationships: MapNet and VISG. Perspectives are those of Phil Wilcox.

MapNet is a genomics collaboration that was formed in 2005 by a collective of New Zealand-based researchers from agriculture, horticulture, forestry and human medical genetics that quickly identified analytical gaps in international statistical genetics research. In response, MapNet members formed the Virtual Institute of Statistical Genetics (i.e. VISG) in 2007 and successfully obtained research funding to address these gaps. Through these synergies, methods developed for large human data sets (e.g. *CNVrd*, *CNVrd2*, *selectionTools*; Nguyen *et al.* 2013, 2014; Cadzow *et al.* 2014) have been successfully applied to apple (*Malus pumila*) data to identify genes of interest in commercial species (e.g. Boocock *et al.* 2015). Other workflows, such as the *selectionTools* pipeline developed and applied to human data sets such as the *1000Genomes* human data (Cadzow *et al.* 2014), are applicable to other outcrossed species where genetic maps are available. Recently, these relationships have also expanded to include cross-sector projects with scientists from the EEB and conservation genetics sector, who are able to provide insight into how these pipelines can be more broadly applicable to other applied genetic disciplines.

Critical for these cross-sector collaborations is effective and ethical behaviours among researchers, distributed leadership, commitment to an explicitly articulated vision, and effective resourcing for method development and testing. Ongoing cost reductions in both high-throughput sequencing and genotyping will constantly challenge data analyses. Thus, collaborations among researchers in primary industry, human medical genetics, EEB and conservation genetics are an effective option to develop and apply genomic methods in a financially limited environment.

The benefits of the above-mentioned collaborations would ensure (a) relevant data analysis tools could be produced by adding relevance and utility to primary-sector researchers proposing to develop such tools, and (b) providing a platform for more efficient utilization of resources such as laboratory spaces and analytical capabilities, further reducing costs and therefore increasing data generation capacity. Collaborating with primary-sector researchers working on closely related species would also benefit conservation genetics by improving efficiency. In some cases, the same species may be endangered within its natural range, but be of commercial value in other regions – such as *Pinus radiata* D. Don, which is widely planted as an exotic in the southern hemisphere but endangered in its natural range in Baja and northern California. An additional benefit of such collaborations is valuable experience and learnings from primary-sector colleagues regarding experiment design, data analyses and interpretation of results. The MapNet collective was formed and run at essentially no cost, by utilizing the resources of collaborating institutions and labour of those who were committed to this initiative, thus such cross-sector networks are easy to establish and operate – and often professionally rewarding for all involved.

provide this biodiversity expertise to commercial species for improved primary production (Fig. 2).

Conservation biologists and primary industry scientists also share similar goals regarding how best to mitigate the impact of climate change (Kristensen *et al.* 2015). For example, plant and animal breeders are prioritizing the selection of heat-tolerant plants (Ye *et al.* 2015) and low-emission animals (Hayes *et al.* 2013) and conservation scientists are debating a role for intentional introgression of desired phenotypic traits (e.g. heat tolerance) among locally adapted species or populations (Hamilton & Miller 2015; Kovach *et al.* 2016; Miller & Hamilton 2016). Given these shared goals, there is merit for scientists in primary industry and conservation to work together to maintain the evolutionary potential of commercial and threatened species in a changing climate.

A compelling rationale for building strong relationships between primary industry and conservation biology is that scientists in both disciplines conduct applied genetic research. Whereas primary industry scientists respond to the needs of primary industry practitioners (i.e. plant and animal breeders, farmers, fishermen and loggers), conservation scientists respond to the needs of conservation practitioners (i.e. wildlife managers and policy makers; Gordon *et al.* 2014; Haig *et al.* 2016). Considering the research–

implementation gap that has been discussed in conservation genetic and genomic literature (Knight *et al.* 2008; Laikre *et al.* 2010; Shafer *et al.* 2015; Taylor & Soanes 2016), researchers from conservation genetics and primary industry can collaborate on how to best communicate research needs and results between scientists and practitioners. In the policy arena, both conservation geneticists and primary industry scientists work to develop improved policy regarding the utilization and dissemination of genetic and genomic information (e.g. the *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization*, <https://www.cbd.int/abs>; the *International Treaty on Plant Genetic Resources for Food and Agriculture*, <http://www.planttreaty.org>) and we anticipate that relationships between the two disciplines will allow for discussion on how to best form policy regarding the application of genomic information to threatened and commercial species.

Cross-sector collaborations will provide exciting opportunities to strategize how best to engage with stakeholders (e.g. private landowners, local governments and research-funding bodies; Jacobson & Duff 1998; Dubbeling & Merzthal 2006); but where we see an even greater opportunity for considerable gains is for conservation geneticists

and primary industry scientists to learn from one another about the importance of building meaningful partnerships with local and indigenous communities. Partnering with these communities enriches conservation and primary industry science because it creates research projects that are informed by the traditional knowledge and needs of these communities from the initial research proposal to the final report. In New Zealand, scientists and practitioners have clear directives to engage with Māori (indigenous peoples of Aotearoa/New Zealand) regarding the management of taonga (treasured) species (i.e. *Ko Aotearoa Tēnei/This is New Zealand*, conventionally known as WAI 262, <http://www.waitangitribunal.govt.nz/>) and various approaches have been developed to facilitate such engagement (Tipene-Matua & Henaghan 2007; Wilcox *et al.* 2008; Hudson *et al.* 2010). In addition, researchers are required to consult with relevant Māori tribes (iwi or hapū) when applying to receive permits for scientific research on taonga species from the Department of Conservation. New Zealand endemic species of cultural importance include threatened species (e.g. tutu-ruatu/shore plover and kākī/black silt; Box 1) and commercial species (e.g. kūtai/green-lipped mussel/Greenshell™ mussel, *Perna canaliculus*), and therefore, we urge conservation genetic and primary industry scientists to collaborate on how to build productive partnerships with relevant Māori communities to develop research that is responsive to the needs and expectations of those communities. Beyond New Zealand, researchers based in any of the 92 countries around the world that are signatories to the *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization* (<https://www.cbd.int/abs/>) have an opportunity to do the same. However, we argue that as global citizens, all scientists should be acting as if their country was a signatory, because as we get closer to generating population genomic data sets that include whole genomes for species of cultural importance we need to be more aware of how these genomic resources can affect and benefit local and indigenous communities.

### Moving forward

While multitasking empirical research, relationships with practitioners, stakeholders and interdisciplinary partnerships can be cumbersome, we are confident that the biggest gains in both conservation genetics and primary industry will be made under this approach. Given the mutual problems that can be solved when conservation geneticists and primary industry scientists work together, we encourage scientists in both disciplines to be leaders in interdisciplinary research and we offer the following advice on how to best forge these relationships:

#### *Get out of your silo*

The first step to building successful interdisciplinary relationships is for researchers to get out of their silos and meet people with aligned research goals across disciplines.

To accomplish this task for conservation genetics and primary industry, we advocate for small (<100 people) and diverse cross-sector meetings that allow participants from academia, government agencies and private institutions to actively engage with every presentation, especially those outside of their silos. In a New Zealand context, annual meetings such as MapNet (see Box 3), the Canterbury 'Omics Symposium and the Queenstown Research Week exemplify small, diverse, cross-sector meetings that allow scientists from both conservation and primary industry to meet and expand their research networks. For larger countries, these diverse and small meetings might be more effective on a regional vs. a national level. In addition to meetings, we encourage conservation geneticists and primary industry scientists to attend genomic and networking workshops to meet people with aligned vision for genomic research, albeit in another discipline.

#### *Practice leadership in interdisciplinary research*

The second step to forging mutually beneficial partnerships between conservation and primary industry is to actively communicate with and collaborate with researchers outside of one's silo. Doing so invariably requires leadership, respect and motivation to tackle shared problems (see Table 1), generally by expanding your own research program to incorporate collaborative interdisciplinary projects between conservation and primary industry (e.g. Banks 2004; Knowler & Bradshaw 2007; Hobbs *et al.* 2008; Blank 2013; Sardinas & Kremen 2015; Box 3). Upon launching these collaborations, it is essential that leaders from both parties open an honest dialog concerning expectations, limitations and potential hindrances to interdisciplinary work such as intellectual property issues. If collaborative groups choose to develop new methods or bioinformatic pipelines, we encourage these groups to test these tools on different species representing a wide range of genomic complexities (i.e. ploidy levels, genome size and number of repetitive elements, see Table 1) so these tools are robust and widely applicable to any research study (see also Boxes 2 and 3). We also advocate for these collaborative groups to develop methods and pipelines that are open source (see Box 2), which inspires others to use and improve upon cross-disciplinary tools. Pursuing co-funding opportunities between conservation and primary industry can be an excellent means of building mutually beneficial research collaborations, especially given that most grant providers favour collaborative proposals that tackle complex problems with broad research impact (Ledford 2015; but see also Bromham *et al.* 2016). Worldwide, there are groups that are forming to tackle complex problems through an interdisciplinary approach, including the Virtual Institute of Statistical Genetics (see Box 3) and Te Pūnaha Matatini (translated to 'the meeting place of many faces', <http://www.tepunahamatatini.ac.nz/>). As leaders from conservation and primary industry initialize interdisciplinary research, we encourage the formation and utilization of these groups to facilitate the scientific process and encourage the involvement of new partners.



*Promote a community of interdisciplinary research*

Leaders in both the conservation and primary industry sphere can go beyond collaborating with interdisciplinary scientists to promote a culture of interdisciplinary research. To accomplish this, we encourage editorial teams at conservation and EEB journals with a broad readership like *Molecular Ecology* to periodically invite perspective articles from colleagues in primary industry. We equate this approach to the recent decision made by the editorial team at *Animal Conservation* to invite submissions from conservation practitioners so conservation academics can better understand the needs and challenges of real-world conservation (Gordon *et al.* 2014). Leaders who are organizing meetings and conferences in primary industry, conservation and genomics can strive to incorporate cross-sector talks and break down organizational silos by minimizing field-specific sessions, as proposed by Taylor & Soanes (2016) and practised by cross-sector meetings like MapNet (see Box 3). We also challenge scientists in both primary industry and conservation to become good interdisciplinary mentors to promote a culture of interdisciplinary research. This can involve mentors in conservation and primary industry promoting genomic seasonal internships or research positions to students in different silos. Not only will this encourage an interdisciplinary field, but it will also produce well-rounded and informed early-career researchers with excellent interpersonal skills and a network of colleagues to help solve shared problems.

After relationships between conservation genetics and primary industry are forged, we do not anticipate relationships will end once genomic gains are made in both disciplines. Instead, we envision these relationships will continue to grow and enable both disciplines to problem-solve and incorporate new technologies for the improvement in threatened and commercial species. With other emerging techniques being discussed and used in both conservation and primary industry, including other -omic techniques (e.g. transcriptomics, proteomics, metabolomics; Diz & Calvete 2016; Todd *et al.* 2016), epigenetic studies (Verhoeven *et al.* 2016) and genome editing (Johnson *et al.* 2016), we expect conservation genetics and primary industry to continue to collaborate and solve mutual problems while incorporating new technologies in applied disciplines.

We are confident that building strong interdisciplinary relationships will enable genomic advances in both conservation genetics and primary industry. However, we appreciate our colleagues in the global conservation community may be pursuing different strategies to successfully navigate the transition from genetics to genomics and we look forward to hearing about them in due course. In the meantime, our hope is that new technologies including genomics will be effectively incorporated into applied genetic disciplines like conservation and primary industry, because there is much to gain using HTS technologies to improve outcomes for the world's threatened and commercial species.

**Acknowledgements**

We appreciate the constructive insights from all anonymous reviewers of this article, particularly in regard to comments that helped us shape the action plan for interdisciplinary research between conservation genetics and primary industry. We thank Margaret Paterson, Mauricio Gonzales Chang and Ilna Cubrinovska for their assistance with the literature search. We are grateful to Paul Gardner for inspiring this manuscript. We acknowledge Marty Faville for his valuable discussion. Finally, we thank Leonie Heyder for the tuturuatu image.

**References**

- Ali OA, O'Rourke SM, Amish SJ *et al.* (2016) RAD capture (rap-ture): flexible and efficient sequence-based genotyping. *Genetics*, **202**, 389–400.
- Allen AM, Barker GL, Wilkinson P *et al.* (2012) Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, **11**, 279–295.
- Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.
- Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.
- Angeloni F, Wagemaker N, Vergeer P, Ouborg J (2011) Genomic toolboxes for conservation biologists. *Evolutionary Applications*, **5**, 130–143.
- Aslam ML, Bastiaansen JWM, Elferink MG *et al.* (2012) Whole genome SNP discovery and analysis of genetic diversity in Turkey (*Meleagris gallopavo*). *BMC Genomics*, **13**, 391.
- Aulchenko S, de Koning D, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577–585.
- Avis JC (2008) The history, purview, and future of conservation genetics. In: *Conservation Biology: Evolution in Action* (eds Carroll SP, Fox CW), pp. 5–15. Oxford University Press, Oxford.
- Avis JC (2010) Perspective: conservation genetics enters the genomics era. *Conservation Genetics*, **11**, 665–669.
- Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.
- Ball RD (2005) Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Genetics*, **170**, 859–873.
- Banks JE (2004) Divided culture: integrating agriculture and conservation biology. *Frontiers in Ecology and the Environment*, **2**, 537–545.
- Bataillon TM, David JL, Schoen DJ (1996) Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, **144**, 409–417.
- Begum H, Spindel JE, Lalusin A *et al.* (2015) Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PLoS ONE*, **10**, e0119873.
- Benestan LM, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, **24**, 3299–3315.

- Benestan LM, Ferchaud L, Hohenlohe PA *et al.* (2016) Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular Ecology*, **25**, 2967–2977.
- Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2016) Genomic analysis reveals depression due to both individual and maternal inbreeding in a free-living mammal population. *Molecular Ecology*, **25**, 3152–3168.
- Bertioli DJ, Ozias-Akins P, Chu Y *et al.* (2014) The use of SNP markers for linkage mapping in diploid and tetraploid peanuts. *G3*, **4**, 89–96.
- Blank PJ (2013) Northern bobwhite response to conservation reserve program habitat and landscape attributes. *The Journal of Wildlife Management*, **77**, 68–74.
- Boocock J, Chagne D, Merriman TR, Black MA (2015) The distribution and impact of common copy-number variation in the genome of the domesticated apple, *Malus × domestica* Borkh. *BMC Genomics*, **16**, 848.
- Bowden R, MacFie TS, Myers S *et al.* (2012) Genomic tools for evolution and conservation in the chimpanzee: *Pan troglodytes ellioti* is a genetically distinct population. *PLoS Genetics*, **8**, e1002504.
- Brachi B, Meyer CG, Villoutreix R *et al.* (2015) Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proceedings of the National Academy of Science*, **112**, 4032–4037.
- Bromham L, Dinnage R, Hua X (2016) Interdisciplinary research has consistently lower funding success. *Nature*, **534**, 684–687.
- Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA (2012) Development and mapping of SNP assays in allotetraploid cotton. *Theoretical Applied Genetics*, **124**, 1201–1214.
- Caballero A, García-Dorado A (2013) Allelic diversity and its implications for the rate of adaptation. *Genetics*, **195**, 1373–1384.
- Cadzow M, Boocock J, Nguyen HT, Wilcox P, Merriman TR, Black MA (2014) A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics*, **5**, 293.
- Card DC, Schield DR, Reyes-Velasco J *et al.* (2014) Two low coverage bird genomes and a comparison of reference-guided versus *de novo* genome assemblies. *PLoS ONE*, **9**, e106649.
- Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.
- Chaves JA, Cooper EA, Hendry AP *et al.* (2016) Genomic variation at the tips of the adaptive radiation of Darwin's finches. *Molecular Ecology*, doi:10.1111/mec.13743.
- Chen N, van Hout CV, Gottipati S, Clari AG (2014) Using mendelian inheritance to improve high-throughput SNP discovery. *Genetics*, **198**, 847–857.
- Clevenger J, Chararro C, Pearl SA, Ozias-Akins P, Jackson SA (2015) Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. *Molecular Plant*, **8**, 831–846.
- Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B*, **363**, 557–572.
- Cubrinovska I, Massaro M, Hale ML (2016) Assessment of hybridisation between the endangered Chatham Island black robin (*Petroica traversi*) and the Chatham Island tomtit (*Petroica macrocephala chathamensis*). *Conservation Genetics*, **17**, 259–265.
- Dai S, Wang C, Wang Z *et al.* (2015) Inbreeding and its effects on fleece traits of inner Mongolia cashmere goats. *Small Ruminant Research*, **128**, 50–53.
- Defaveri J, Viitaniemi H, Leder E, Merilä J (2013) Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. *Molecular Ecology Resources*, **13**, 377–392.
- Dierickx EG, Shultz AJ, Sato F, Hiraoka T, Edwards SV (2015) Morphological and genomic comparisons of Hawaiian and Japanese black-footed albatrosses (*Phoebastria nigripes*) using double digest RADseq: implications for conservation. *Evolutionary Applications*, **8**, 662–678.
- Dillon SK, Nolan MF, Matter P, Gapare WJ, Bragg JG, Southerton SG (2013) Signatures of adaptation and genetic structure among the mainland populations of *Pinus radiata* (D. Don) inferred from SNP loci. *Tree Genetics & Genomes*, **9**, 1447–1463.
- Diz AP, Calvete JJ (2016) Ecological proteomics: is the field ripe for integrating proteomics into evolutionary ecology research? *Journal of Proteomics*, **135**, 1–3.
- Dodds KG, McEwan JC, Brauning R *et al.* (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, **16**, 1047.
- Dowling TE, Anderson CD, Marsh PC, Rosenberg MS (2015) Population structure in the roundtail chub (*Gila robusta* complex) of the Gila River Basin as determined by microsatellites: evolutionary and conservation implications. *PLoS ONE*, **10**, e0139832.
- Dubbeling M, Merzthal G (2006) Sustaining urban agriculture requires the involvement of multiple stakeholders. In: *Cities Farming for Future, Urban Agriculture for Green and Productive cities* (ed. van Veenhuizen R), pp. 20–51. IDRC and IIRR Publishing, Ottawa, CA.
- Edmonds S (2007) Between a rock and a hard place: evaluating the relative risks of inbreeding and outbreeding depression for conservation and management. *Molecular Ecology*, **16**, 463–475.
- Eklom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
- Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, **29**, 51–63.
- Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.
- Eteqadi B, Hossein-Zadeh NG, Shadparvar AA (2015) Inbreeding effects on reproductive traits in Iranian Guilan sheep. *Tropical Animal Health and Production*, **47**, 533–539.
- Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*, 4th edn. Longman, Essex.
- Frankham R (1995) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.
- Frankham R (2005) Genetics and extinction. *Biological Conservation*, **126**, 131–140.
- Frankham R (2010a) Where are we in conservation genetics and where do we need to go? *Conservation Genetics*, **11**, 661–663.
- Frankham R (2010b) Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, **143**, 1919–1927.
- Frankham R (2015) Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Molecular Ecology*, **24**, 2610–2618.
- Frankham R, Ballou JD, Eldridge MD *et al.* (2011) Predicting the probability of outbreeding depression. *Conservation Biology*, **25**, 265–275.
- Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology and Evolution*, **27**, 489–496.
- Garner BA, Hand BK, Amish SJ *et al.* (2016) Genomics in conservation: case studies and bridging the gap between data and application. *Trends in Ecology and Evolution*, **31**, 81–83.
- George AW, Visscher PM, Haley CS (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, **156**, 2081–2092.

- Glaubitx JC, Casstevens TM, Lu F *et al.* (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, **9**, e90346.
- Goff SA, Ricke D, Lan TH *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. Japonica). *Science*, **296**, 92–100.
- Gordon IJ, Evans DM, Garnder TWJ *et al.* (2014) Enhancing communication between conservation biologists and conservation practitioners: letter from the conservation front line. *Animal Conservation*, **17**, 1–2.
- Gore MA, Chia JM, Elshire RJ *et al.* (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115–1117.
- Groom MJ, Meffe GK, Carroll CR (2006). *Principles of Conservation Biology* (No. 333.9516G7). Sinauer Associates, Sunderland.
- Grueber CE (2016) Comparative genomics for biodiversity conservation. *Computational and Structural Biotechnology*, **13**, 370–375.
- Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nature*, **37**, 549–554.
- van Gurp TP, Wagemaker NC, Wouters B, Vergeer P, Ouborg JN, Verhoeven KJ (2016) epiGBS: reference-free reduced representation bisulfite sequencing. *Nature Methods*, **13**, 322–324.
- Gutiérrez JP, Marmi J, Goyache F, Jordana J (2005) Pedigree information reveals moderate to high levels of inbreeding and a weak population structure in the endangered Catalanian donkey breed. *Journal of Animal Breed Genetics*, **122**, 378–386.
- Haas RJ, Payseur BA (2016) Fifteen years of genomewide scans for selection: trends, lessons, and unaddressed genetic sources of complication. *Molecular Ecology*, **25**, 5–23.
- Hagen EN, Hale ML, Maloney RF, Steeves TE (2011) Conservation genetic management of a critically endangered New Zealand endemic bird: minimizing inbreeding in the black stilt *Himantopus novaezelandiae*. *Ibis*, **153**, 556–561.
- Haig SM, Miller MP, Bellinger R, Draheim HM, Mercer DM, Mullins TD (2016) The conservation genetics juggling act: integrating genetics and ecology, science and policy. *Evolutionary Applications*, **9**, 181–195.
- Halbert ND, Ward TJ, Schnabel RD, Taylor JF, Derr JN (2005) Conservation genomics: disequilibrium mapping of domestic cattle chromosomal segments in North American bison populations. *Molecular Ecology*, **14**, 2343–2362.
- Hamilton JA, Miller JM (2015) Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology*, **30**, 33–41.
- Hammerly SC, de la Cerda DA, Bailey J, Johnson JA (2016) A pedigree gone bad: increased offspring survival after using DNA-based relatedness to minimize inbreeding in a captive population. *Animal Conservation*, **19**, 296–303.
- Hapke A, Thiele D (2016) GlibPSs: a toolkit for fast and accurate analyses of genotyping-by-sequencing data without a reference genome. *Molecular Ecology Resources*, **16**, 979–990.
- Harrison KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014) Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, **7**, 1008–1025.
- Hayes BJ (2011) Technical note: efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. *Journal of Dairy Science*, **94**, 2114–2117.
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science*, **92**, 433–443.
- Hayes BJ, Lewin JA, Goddard ME (2013) The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends in Genetics*, **29**, 206–214.
- He J, Zhao X, Laroche A, Lu Z, Liu H, Li Z (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, **5**, 484.
- He X, Johansson ML, Heath DD (2016) Role of genomics and transcriptomics in selection of reintroduction source populations. *Conservation Biology*, **30**, 1010–1018.
- Heffner EL, Sorrells ME, Jannink J (2009) Genomic selection for crop improvement. *Crop Science*, **49**, 1–12.
- Herrero-Medrano JM, Megens HJ, Crooijmans RP, Abellaneda JM, Ramis G (2012) Farm-by-farm analysis of microsatellite, mtDNA, and SNP genotype data reveals inbreeding and crossbreeding as threats to the survival of a native Spanish pig breed. *Animal Genetics*, **44**, 259–266.
- Herten K, Hestand MS, Vermeesch JR, Van Houdt KJ (2015) GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics*, **16**, 73.
- Hess JE, Zandt JS, Matala AR, Narum SR (2016) Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. *Proceedings of the Royal Society B*, **283**, doi:10.1098/rspb.2015/3064.
- Hillier L, Miller W, Birney E *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.
- Hobbs PR, Sayre K, Gupta R (2008) The role of conservation agriculture in sustainable agriculture. *Philosophical Transactions of the Royal Society Biology*, **363**, 543–555.
- Hoffman JL, Simpson F, David P *et al.* (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Science*, **111**, 3775–3780.
- Hogg C, Grueber CE, Pemberton D *et al.* (2016) “Devil Tools & Tech”: a synergy of conservation research and management practices. *Conservation Letters*, doi:10.1111/conl.12221.
- Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.
- Hu ZL, Park CA, Wu XL, Reecy HM (2013) Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Research*, **41**, 871–879.
- Hudson M, Milne M, Reynolds P, Russell K, Smith B (2010) Te Ara Tika Guidelines for Māori research ethics: a framework for researchers and ethics committee members. ISBN: 978-1-877495-03-8.
- IUCN/SSC (2013) *Guidelines for Reintroductions and Other Conservation Translocations. Version 1.0*. IUCN Species Survival Commission, Gland Switzerland; Cambridge, UK.
- Jacobson SK, Duff MD (1998) Training idiot savants: the lack of human dimensions in conservation biology. *Conservation Biology*, **12**, 263–267.
- Jiménez-Mena B, Hospital F, Bataillon T (2016) Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. *Conservation Genetics Resources*, **8**, 35–41.
- Johnson JA, Altwegg R, Evans DM *et al.* (2016) Is there a future for genome-editing technologies in conservation? *Animal Conservation*, **19**, 97–101.
- Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.
- Joost S, Colli L, Bonin A *et al.* (2011) Promoting collaboration between livestock and wildlife conservation genetic communities. *Conservation Genetic Resources*, **3**, 785–788.
- Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, **115**, 63–72.

- Keller LF, Arcese P, Smith JNM, Hochachka WM, Stearns SC (1994) Selection against inbred song sparrows during a natural population bottleneck. *Nature*, **372**, 356–357.
- Kim HB, Joa JH, Song KJ (2015) Current status and prospects of kiwifruit (*Actinidia chinensis*) genomics. *Journal of Plant Biotechnology*, **42**, 342–349.
- Knief U, Hemmrich-Stanisak G, Wittig M *et al.* (2015) Quantifying realized inbreeding in wild and captive animal populations. *Heredity*, **114**, 397–403.
- Knight AT, Cowling RM, Rouget M, Balmford A, Lombard AT, Campbell BM (2008) Knowing but not doing: selecting priority conservation areas and the research-implementation gap. *Conservation Biology*, **22**, 610–617.
- Knowler D, Bradshaw B (2007) Farmer's adoption of conservation agriculture: a review and synthesis of recent research. *Food Policy*, **32**, 25–48.
- Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. *Trends in Ecology and Evolution*, **21**, 629–637.
- Kovach RP, Luikart G, Lower WH, Boyer MC, Muhlfeld CC (2016) Risk and efficacy of human-enabled interspecific hybridization for climate-change adaptation: response to Hamilton and Miller (2016). *Conservation Biology*, **30**, 428–430.
- Kraus RH, Vonholdt B, Cocchiararo B *et al.* (2015) A single-nucleotide polymorphism-based approach for rapid and cost-effective genetic wolf monitoring in Europe based on noninvasively collected samples. *Molecular Ecology Resources*, **15**, 295–305.
- Kristensen TN, Hoffmann AA, Pertoldi C, Stronen AV (2015) What can livestock breeders learn from conservation genetics and vice versa? *Frontiers in Genetics*, **6**, 1–12.
- Laike L, Allendorf FW, Aroner LC *et al.* (2010) Neglect of genetic diversity in implementation of the convention on biological diversity. *Conservation Biology*, **24**, 86–88.
- Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, **7**, 355–369.
- Ledford H (2015) Team science. *Nature*, **525**, 308–311.
- Lemay MA, Russello MA (2015) Genetic evidence for ecological divergence in kokanee salmon. *Molecular Ecology*, **24**, 798–811.
- Lenstra JA, Groeneveld LF, Eding J *et al.* (2012) Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. *Animal Genetics*, **43**, 483–502.
- Leroy G, Mary-Huard T, Danvy S, Charvolin E, Danchin-Burge C (2013) Methods to estimate effective population size using pedigree data: examples in dog, sheep, cattle, and horse. *Genetics Selection Evolution*, **45**, 1.
- Lew RM, Finger AJ, Baerwald MR, Goodbla A, May B, Meek MH (2015) Using next-generation sequencing to assist a conservation hatchery: a single-nucleotide polymorphism panel for the genetic management of endangered delta smelt. *Transactions of the American Fisheries Society*, **144**, 767–779.
- Li M, Strandén I, Tiirikka T, Sevón-Aimonen M, Kantanen J (2011) A comparison of approaches to estimate the inbreeding coefficient and pairwise relatedness using genomic and pedigree data in a sheep population. *PLoS ONE*, **6**, e26256.
- Li S, Li B, Cheng C *et al.* (2014) Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biology*, **15**, 557.
- Li Y, Lancaster ML, Cooper SJB, Taylor AC, Carthew SM (2015a) Population structure and gene flow in the endangered southern brown bandicoot (*Isodon obesulus obesulus*) across a fragmented landscape. *Conservation Genetics*, **16**, 331–345.
- Li X, Buitenhuis AJ, Lund MS *et al.* (2015b) Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. *Journal of Dairy Science*, **98**, 8152–8163.
- Ljungqvist M, Åkesson M, Hansson B (2010) Do microsatellites reflect genome-wide genetic diversity in natural populations? A comment on Väli *et al.* (2008). *Molecular Ecology*, **19**, 851–855.
- Lu F, Lipka AE, Glaubitz J *et al.* (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genetics*, **9**, e1003215.
- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.
- Luković Z, Karolyi D, Klišanić V, Mahnet Ž, Škorput D (2012) Genetic parameters and trends for litter size in Black Slavonian pigs. 7th International Symposium on the Mediterranean Pig, ISBN: 2-85352-488-4.
- Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland.
- Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Marshall C, Chagné D, Deusch O *et al.* (2015) A DNA-based diagnostic for differentiating among New Zealand endemic *Podocarpus*. *Tree Genetics & Genomes*, **11**, 1–13.
- Matukumalli LK, Lawley CT, Schnabel RD *et al.* (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE*, **4**, e5350.
- McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.
- McMahon BJ, Teeling EC, Höglund J (2014) How and why should we implement genomics into conservation? *Evolutionary Applications*, **7**, 999–1007.
- Medugorac I, Medugorac A, Russ I *et al.* (2009) Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. *Molecular Ecology*, **18**, 3394–3410.
- Melo ATO, Bartaula R, Hale I (2016) GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics*, **17**, 29.
- Miller JM, Hamilton JA (2016) Interspecies hybridization in the conservation toolbox: response to Kovach *et al.* 2016. *Conservation Biology*, **30**, 431–433.
- Miller JM, Poissant J, Hogg JT, Coltman DW (2012) Genomic consequences of genetic rescue in an insular population of bighorn sheep (*Ovis canadensis*). *Molecular Ecology*, **21**, 1583–1596.
- Moula N, Antoine-Moussiaux N, Farnir F, Leroy P (2009) Evaluation of the production performances of an endangered local poultry breed, the famennoise. *International Journal of Poultry Science*, **8**, 389–396.
- Muir WM, Wong GKS, Zhang Y *et al.* (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Science*, **105**, 17312–17317.
- Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.
- Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, **12**, 111–112.
- Nguyen HT, Merrimen TR, Black MA (2013) CNVrd, a read-depth algorithm for assigning copy-number at the FCGR locus:



- population-specific tagging of copy number variation at FCGR3B. *PLoS ONE*, **8**, e63219.
- Nguyen HT, Merrimen TR, Black MA (2014) The CNVrd2 package: measurement of copy number at complex loci using high-throughput sequencing data. *Frontiers in Genetics*, **5**, 248.
- Ogden R, Baird J, Senn H, McEwing R (2012) The use of cross-species genome-wide arrays to discover SNP markers for conservation genetics: a case study from Arabian and scimitar-horned oryx. *Conservation Genetic Resources*, **4**, 471–473.
- Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics*, **26**, 177–187.
- Overbeek AL, Hauber ME, Brown E, Cleland S, Maloney RF, Steeves TE (2016) Evidence for brood parasitism in a critically endangered Charadriiform with implications for conservation. *Journal of Ornithology*, doi:10.1007/s/0336-016-1375-x.
- Oyler-McCance SJ, Cornman RS, Jones KL, Fike JA (2015) Genomic single-nucleotide polymorphisms confirm that Gunnison and greater sage-grouse are genetically well differentiated and that the bi-state population is distinct. *Condor*, **117**, 217–227.
- Pacioni C, Hunt H, Allentoft ME *et al.* (2015) Genetic diversity loss in a biodiversity hotspot: ancient DNA quantifies genetic decline and former connectivity in a critically endangered marsupial. *Molecular Ecology*, **24**, 5813–5828.
- Pardo-Diaz C, Salazar C, Jiggins CD (2015) Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, **6**, 445–464.
- Payseur BA, Rieseberg LH (2016) A genomic perspective on hybridization and speciation. *Molecular Ecology*, **25**, 2337–2360.
- Pertoldi C, Wójcik JM, Tokarska M *et al.* (2009) Genome variability in European and American bison detected using the BovineSNP50 bead chip. *Conservation Genetics*, **11**, 627–634.
- Poissant J, Shafer BA, Davis S *et al.* (2009) Genome-wide cross-amplification of domestic sheep microsatellites in bighorn sheep and mountain goats. *Molecular Ecology Resources*, **9**, 1121–1126.
- Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**, 459–463.
- Primmer CR (2009) From conservation genetics to conservation genomics. *Annals of the New York Academy of Science*, **1162**, 357–368.
- Prince DJ, O'Rourke SM, Thompson TQ *et al.* (2016) The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *bioRxiv*, doi:10.1101/056853.
- Pryce JE, Hayes BJ, Goodard ME (2012) Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *Journal of Dairy Science*, **95**, 377–388.
- Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RAD seq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, **2**, e431.
- Rietveld CA, Medland SE, Derringer J *et al.* (2013) GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, **340**, 1467–1471.
- Robinson JA, Vecchyo DO, Fan Z (2016) Genomic flatlining in the endangered island fox. *Current Biology*, **26**, 1183–1189.
- Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.
- Sardinas HX, Kremen C (2015) Pollination services from field-scale agricultural diversification may be context-dependent. *Agriculture, Ecosystems and Environment*, **207**, 17–25.
- Saura M, Fernández A, Varona L *et al.* (2015) Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data. *Genetics Selection Evolution*, **47**, 1.
- Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.
- Schibler L, Roig A, Mahe MF *et al.* (2004) A first generation bovine BAC-based physical map. *Genetics selection evolution*, **36**, 105–122.
- Schielzeth H, Husby A (2014) Challenges and prospects in genome-wide quantitative trait loci mapping of standing genetic variation in natural populations. *Annals of the New York Academy of Sciences*, **1320**, 35–57.
- Seddon PJ (2010) From reintroduction to assisted colonization: moving along the conservation translocation spectrum. *Restoration Ecology*, **18**, 796–802.
- Sellner EM, Kim JW, McClure MC, Taylor KH, Schnabel RD, Taylor JF (2007) Board-invited review: applications of genomic information in livestock. *Journal of Animal Science*, **85**, 3148–3158.
- Shafer ABA, Wolf JBW, Alves PC *et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, **30**, 78–87.
- Shafer ABA, Wolf JBW, Alves PC *et al.* (2016) Reply to Garner *et al.* *Trends in Ecology and Evolution*, **31**, 83–84.
- Shepherd LD, de Lange PJ, Cox S, McLenachan PA, Roskrug NR, Lockhart PJ (2016) Evidence of a strong domestication bottleneck in the recently cultivated New Zealand endemic root crop, *Arthropodium cirratum* (Asparagaceae). *PLoS ONE*, **11**, e0152455.
- Siló L, Barragán C, Fernández AI, García-Casco J, Rodríguez MC (2016) Assessing effective population size, coancestry and inbreeding effects on litter size using the pedigree and SNP data in closed lines of the Iberian pig breed. *Journal of Animal Breeding and Genetics*, **133**, 145–154.
- da Silva AG, Appleyard SA, Upston J (2015) Establishing the evolutionary compatibility of potential sources of colonizers for overfished stocks: a population genomics approach. *Molecular Ecology*, **24**, 564–579.
- Slate J, Santure AW, Feulner PGD *et al.* (2010) Genome mapping in intensively studied wild vertebrate populations. *Trends in Ecology and Evolution*, **26**, 275–284.
- Sonah H, Deshmukh RK, Singh VP, Gupta DK, Singh NK, Sharma TR (2011) Genomic resources in horticultural crops: Status, utility and challenges. *Biotechnology Advances*, **29**, 199–209.
- Soulé ME (1985) What is conservation biology? A new synthetic discipline addresses the dynamics and problems of perturbed species, communities, and ecosystems. *BioScience*, **35**, 727–734.
- Sovic MG, Fries AC, Gibbs HL (2015) AfterRAD: a pipeline for accurate and efficient *de novo* assembly of RADseq data. *Molecular Ecology Resources*, **15**, 1163–1171.
- Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology and Evolution*, **25**, 705–712.
- Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL (2006) Whole-genome genotyping with the single-base extension assay. *Nature Methods*, **3**, 31–33.
- Steeves TE, Maloney RF, Hale ML, Tylianakis JM, Gemmell NJ (2010) Genetic analyses reveal hybridization but no hybrid swarm in one of the world's rarest birds. *Molecular Ecology*, **19**, 5090–5100.
- Steiner CC, Putnam AS, Hoeck PEA, Ryder OA (2013) Conservation genomics of threatened animal species. *Annual Review of Animal Bioscience*, **1**, 261–281.
- Taylor HR, Soanes K (2016) Breaking out of the echo chamber: missed opportunities for genetics at conservation conferences. *Biodiversity and Conservation*, **25**, 1987.
- Taylor HR, Kardos MD, Ramstad KM, Allendorf FW (2015) Valid estimates of individual inbreeding coefficients from

- marker-based pedigrees are not feasible in wild populations with low allelic diversity. *Conservation Genetics*, **16**, 901–913.
- Tigano A, Friesen VL (2016) Genomics of local adaptation with gene flow. *Molecular Ecology*, **25**, 2144–2164.
- Tipene-Matua B, Henaghan M (2007) Establishing a Māori ethical framework for genetic research with Māori. In: *Genes, Society and the Future* (ed. Henaghan M), pp. 1–44. Human Genome Research Project, Dunedin, New Zealand.
- Todd EV, Black MA, Gemmell NJ (2016) The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology*, **25**, 1224–1241.
- Trask AE, Bignal EM, McCracken DI, Monaghan P, Piernsey SB, Reid JM (2015) Evidence of the phenotypic expression of a lethal recessive allele under inbreeding in a wild population of conservation concern. *Journal of Animal Ecology*, **85**, 879–891.
- Tuberosa R, Salvi S (2006) Genomics-based approaches to improve drought tolerance of crops. *Trends in Plant Science*, **11**, 405–412.
- Uitendwilligen JGAML, Wolters AA, D'hoop BB, Borm TJA, Visser RGF, VanEck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS*, **10**, e0141940.
- Väli U, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, **17**, 3808–3817.
- Varshney RK, Ribaut J, Buckler ES, Tuberosa R, Rafalski JA, Langridge P (2012) Can genomics boost productivity of orphan crops? *Nature Biotechnology*, **30**, 1172–1176.
- Verhoeven KJF, vonHoldt BM, Sork VL (2016) Epigenetics in ecology and evolution: what we know and what we need to know. *Molecular Ecology*, **25**, 1631–1638.
- Vilas A, Pérez-Figueroa A, Quesada H, Caballero A (2015) Allelic diversity for neutral markers retains a higher adaptive potential for quantitative traits than expected heterozygosity. *Molecular Ecology*, **24**, 4419–4432.
- Wang B, Ekblom R, Bunikis I, Siitari H, Höglund J (2014) Whole genome sequencing of the black grouse (*Tetrao tetrix*): reference guided assembly suggests faster-Z and MHC evolution. *BMC Genomics*, **15**, 180.
- Wang L, Zu L, Liu X (2015) Copy number variation-based genome wide association study reveals additional variants contributing to meat quality in swine. *Scientific Reports*, **5**, 12535.
- Weeks AJ, Sgro CM, Young AG *et al.* (2011) Assessing the benefits and risks of translocations in changing environments: a genetic perspective. *Evolutionary Applications*, **4**, 709–725.
- White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. *Molecular Ecology*, **22**, 2971–2985.
- Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA (2015) Genetic rescue to the rescue. *Trends in Ecology & Evolution*, **30**, 42–49.
- Wilcox PL, Charity JA, Roberts MR *et al.* (2008) A values-based process for cross-cultural dialogue between scientists and Māori. *Journal of the Royal Society of New Zealand*, **38**, 215–227.
- Willoughby JR, Fernandez NB, Lamb MC, Ivy JA, Lacy RC, DeWoody JA (2015) The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Molecular Ecology*, **24**, 98–110.
- Windig JJ, Engelsma KA (2010) Perspectives of genomics for genetic conservation of livestock. *Conservation Genetics*, **11**, 635–641.
- Womack JE (2005) Advances in livestock genomics: opening the barn door. *Genome Research*, **15**, 1699–1705.
- Wuerschum T, Boeven PHG, Langer SM, Longin CFJ, Leiser W (2015) Multiply to conquer: copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genetics*, **16**, 96.
- Xia Q, Zhou S, Lu C *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.
- Ye C, Tenorio F, Argayoso MA *et al.* (2015) Identifying and confirming quantitative trait loci associated with heat tolerance at flowering stage in different rice populations. *BMC Genetics*, **16**, 41.
- Zhang G, Li C, Li Q *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311–1320.
- Zhang C, Wang Z, Bruce H *et al.* (2015) Genome-wide association studies (GWAS) identify a QTL close to PRKAG3 affecting meat pH and colour in crossbred commercial pigs. *BMC Genetics*, **16**, 33.
- Zhou Z, Jiang Y, Wang Z *et al.* (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, **33**, 408–413.

---

S.J.G. and T.E.S. were the lead investigators on this research. S.J.G. and T.E.S. designed the research in collaboration with all authors, S.J.G. compiled the literature search, T.E.S., R.E., and P.W. contributed perspective boxes, and all authors wrote the manuscript and provided feedback on the reviewer's comments.

---

doi: 10.1111/mec.13837

#### Data accessibility

Text files including all literature search results presented in Fig. 1 and Fig. S1 (Supporting information) are available on Dryad: doi: 10.5061/dryad.32j55.

#### Supporting information

Additional supporting information may be found in the online version of this article.

**Data S1** An ISI Web of Science literature search was conducted to compare the number of publications from 2005 to 2015 that have been produced in the fields of conservation biology and primary industry. It should be noted that the number of publications may be underestimated in both disciplines, given that conservation and primary industry researchers can publish their results in the grey literature (Garner *et al.* 2016).

**Fig. S1** Growth in the number of genomic publications utilising high-throughput sequencing from 2005 to 2015 in the fields of primary industry (A) and conservation (B).

## Appendix C: Chapter 3, Supplemental Materials

### C1: Literature Search

An ISI Web of Science literature search was conducted to compare the number of publications from 2005-2015 that have been produced in the fields of conservation biology and primary industry. It should be noted that the number of publications may be underestimated in both disciplines, given that conservation and primary industry researchers often publish their results in the grey literature (Garner *et al.* 2016).

#### C1.1: Conservation Genomics

Conservation genomics papers were searched for using the ISI Web of Science (Advanced Search option, with all databases) and words specific to conservation and high throughput sequencing. The following string was used for this search:  
TS=((((((((((((((((((((("conservation biology" OR ecology) OR "environmental science") OR wildlife) OR biodiversity) OR ecosystem) OR "natural resources") OR "restoration ecology") OR habitat) OR extinct "endangered species") OR "threatened species") OR "small population") OR dispersal) OR migration) OR demographic) OR inbreeding) OR bottleneck) OR hybrid) OR translocation) OR "adaptive variation") OR "cryptic species") OR kinship) OR relatedness) OR pedigree) AND (((recommend OR policy) OR management) OR implication)) AND (((((((((((((((((((genome OR genomics) OR "next-generation sequencing") OR "high-throughput sequencing") OR ngs) OR "genome-wide") OR "single nucleotide polymorphism") OR "differential expression") OR "causal variant") OR gwas) OR "rad sequencing") OR "gbs") OR snp) OR illumina) OR "454") OR "Ion Torrent") OR "snp chip"))

This search yielded approximately 6,423 articles, which were refined to 3,638 articles once duplicates were removed. These were downloaded into EndNote X7 and were further refined to 51 articles as follows: All articles that did not have clear implications for the conservation management of focal threatened plant or animal species were removed including all articles pertaining to environmental DNA and/or metagenomics. All review, perspective and opinion articles were removed. All articles using high-throughput sequencing reads for the sole purpose of microsatellite marker discovery were removed. All articles relating to human health and primary industry were removed. All articles that did not use high-throughput sequencing to generate genomic data were removed including all articles pertaining to transcriptomic and proteomics.

#### C1.2: Primary Industry Genomics

Primary Industry genomics papers were searched for using the ISI Web of Science (Advanced Search option, with all databases) and words specific to primary industry and high-throughput sequencing. The following string was used for this search:

TS=((((((((((((((((((((agriculture OR horticulture) OR food) OR crop) OR fruit) OR vegetable) OR poultry) OR cattle) OR sheep) OR livestock) OR pig) OR "commercial fisheries") OR aquaculture) OR forestry) OR cotton) OR wool) OR hemp) OR silk) OR flax) OR cultivar) AND (((((((((((((((((((((genome OR genomics) OR "next-generation sequencing") OR "high-throughput sequencing") OR ngs) OR "genome-wide") OR "single-nucleotide polymorphism") OR snp) OR illumina) OR "454") OR "ion torrent") OR "snp chip") OR "differential expression") OR "causal variant") OR "rad sequencing") OR gbs) OR "genomic selection") OR gwas))

This search yielded approximately 81,957 articles, which was refined to 39,603 articles once duplicates were removed. These articles were downloaded into EndNote X7 and refined to 1,981 articles as follows: All articles that did not have clear implications for the enhancement of primary production for focal plant or animal species was removed including all articles pertaining to environmental DNA and/or metagenomics. All review, perspective, and opinion articles were removed. All articles relating to human health, ecology and evolution were removed including articles pertaining to food poisoning, swine flue or avian influenza. All articles that did not use high-throughput sequencing to generate genomic data were removed including all articles pertaining to transcriptomic and proteomics.

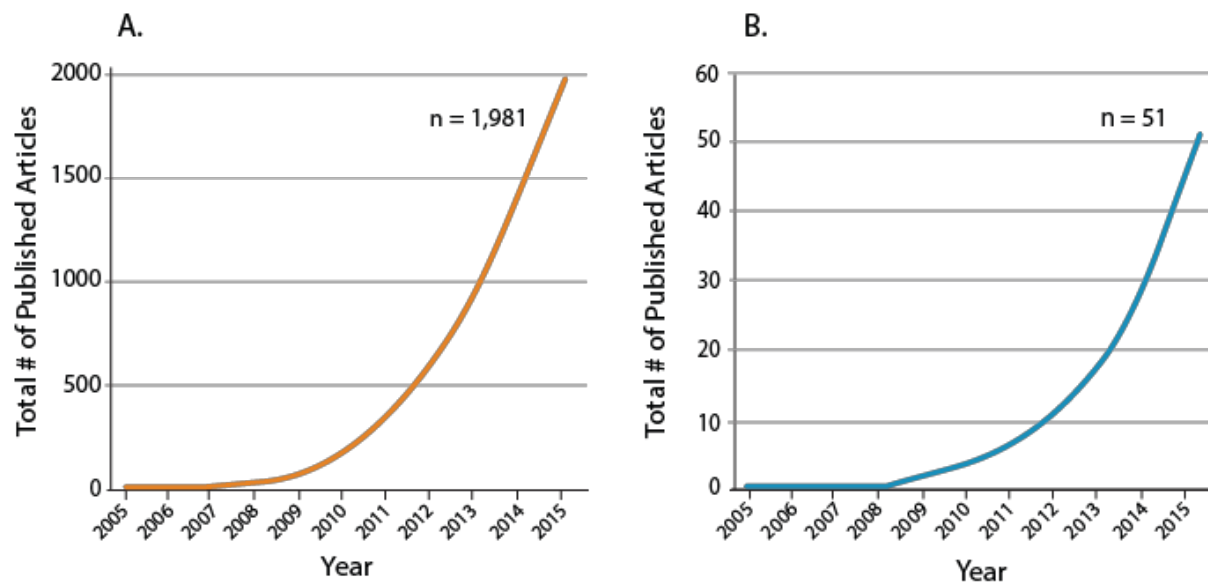


Figure C 1 Growth in the number of genomic publications utilising high-throughput sequencing from 2005-2015 in the fields of primary industry (A) and conservation (B). Curved lines have been smoothed for ease of interpretation.



# Appendix D: Chapter 4, Published Manuscript



Article

## Reference Genomes from Distantly Related Species Can Be Used for Discovery of Single Nucleotide Polymorphisms to Inform Conservation Management

Stephanie J. Galla <sup>1,\*</sup>, Natalie J. Forsdick <sup>2,\*</sup>, Liz Brown <sup>3</sup>, Marc P. Hoepfner <sup>4</sup>, Michael Knapp <sup>2</sup>, Richard F. Maloney <sup>5</sup>, Roger Moraga <sup>6</sup>, Anna W. Santure <sup>7</sup> and Tammy E. Steeves <sup>1</sup>

<sup>1</sup> School of Biological Sciences, University of Canterbury, Christchurch 8140, New Zealand; tammy.steeves@canterbury.ac.nz

<sup>2</sup> Department of Anatomy, University of Otago, Dunedin 9054, New Zealand; michael.knapp@otago.ac.nz

<sup>3</sup> Te Manahuna, Department of Conservation, Twizel 7901, New Zealand; lbrown@doc.govt.nz

<sup>4</sup> Institute of Clinical Molecular Biology, Kiel University, 24105 Kiel, Germany; m.hoepfner@ikmb.uni-kiel.de

<sup>5</sup> Science and Policy Group, Department of Conservation, Christchurch 8011, New Zealand; rmaloney@doc.govt.nz

<sup>6</sup> Tea Break Bioinformatics Ltd., Palmerston North 4144, New Zealand; r.moraga@protonmail.com

<sup>7</sup> School of Biological Sciences, University of Auckland, Auckland 1142, New Zealand; a.santure@auckland.ac.nz

\* Correspondence: stephanie.galla@pg.canterbury.ac.nz (S.J.G.); natalie.forsdick@postgrad.otago.ac.nz (N.J.F.)

Received: 8 November 2018; Accepted: 19 December 2018; Published: 22 December 2018



**Abstract:** Threatened species recovery programmes benefit from incorporating genomic data into conservation management strategies to enhance species recovery. However, a lack of readily available genomic resources, including conspecific reference genomes, often limits the inclusion of genomic data. Here, we investigate the utility of closely related high-quality reference genomes for single nucleotide polymorphism (SNP) discovery using the critically endangered kakī/black stilt (*Himantopus novaezelandiae*) and four Charadriiform reference genomes as proof of concept. We compare diversity estimates (i.e., nucleotide diversity, individual heterozygosity, and relatedness) based on kakī SNPs discovered from genotyping-by-sequencing and whole genome resequencing reads mapped to conordinal (killdeer, *Charadrius vociferus*), confamilial (pied avocet, *Recurvirostra avosetta*), congeneric (pied stilt, *Himantopus himantopus*) and conspecific reference genomes. Results indicate that diversity estimates calculated from SNPs discovered using closely related reference genomes correlate significantly with estimates calculated from SNPs discovered using a conspecific genome. Congeneric and confamilial references provide higher correlations and more similar measures of nucleotide diversity, individual heterozygosity, and relatedness. While conspecific genomes may be necessary to address other questions in conservation, SNP discovery using high-quality reference genomes of closely related species is a cost-effective approach for estimating diversity measures in threatened species.

**Keywords:** conservation genomics; conservation genomics gap; SNP discovery; B10K; threatened species; birds

### 1. Introduction

The field of conservation genetics is in transition from using relatively few genetic markers (e.g., microsatellites, mitochondrial sequences) to using thousands of genome-wide single nucleotide polymorphisms (SNPs) discovered with high-throughput sequencing technologies (HTS) to inform conservation management of threatened species. In addition to providing greater resolution for

diversity estimates (e.g., nucleotide diversity, heterozygosity, relatedness [1]), these new genomic approaches provide an opportunity to tackle new questions regarding regions of the genome that underlie fitness-related traits (i.e., adaptive variation [2–4]). While the promise of a conservation genomic approach has been heralded for well over a decade [5], the uptake of these technologies by conservation management has been slow [6,7].

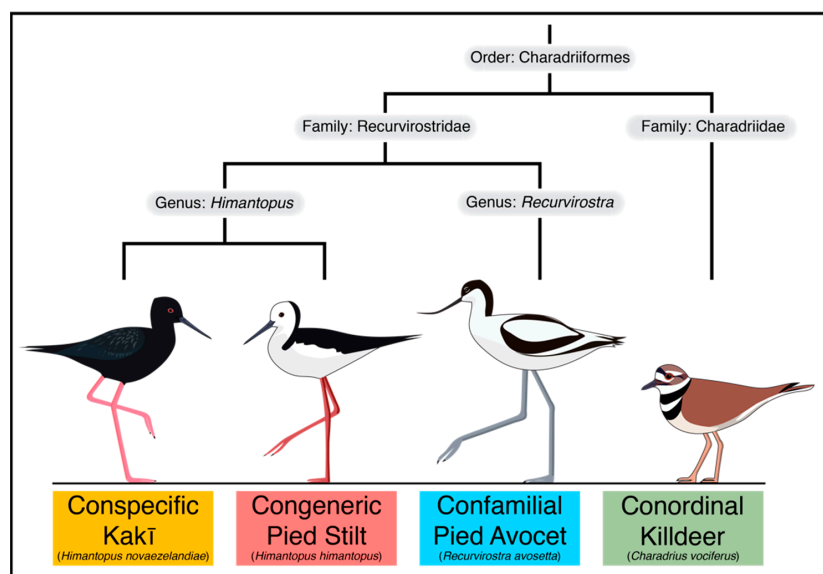
This time lag between technology availability and implementation (also termed the ‘conservation genomics gap’ [7]) may be caused by several interconnected issues, including a disconnect between conservation genetic researchers and practitioners [8,9], the time it takes for geneticists to upskill in bioinformatic expertise [6,7,10], and initial expense for HTS sequence production and generation of genomic resources (e.g., a high-quality reference genome). With that said, sequencing costs are dropping precipitously [11] (but see also [12]) and affordable reduced representation genomic approaches provide the ability to produce high-density marker sets, even in the absence of a reference genome (i.e., de novo marker discovery [13]). While it is possible to discover SNPs de novo, reference-guided approaches to SNP discovery offer many advantages, including enhanced computational efficiency, improved accuracy at low sequencing depth, higher confidence in identifying sequence contamination, greater ability to identify the location of SNPs, improved performance in determining linkage disequilibrium between SNPs, and greater ability to identify differences between paralogous and repetitive sequences from true SNP variants [14–17]. Reference genomes also allow for identifying variants in annotated gene regions, which is necessary for identifying adaptive variation [14]. While reference genomes are preferred for conservation genomic research, they are often unavailable for threatened species or out of reach for resource-constrained conservation projects (e.g., [18]).

There has been an exponential increase in the number of available eukaryotic genomes for non-model species that may be used as a reference [19], including the outputs from various genome consortiums (e.g., Genome 10K [20]; Bird 10,000 Genomes Project (B10K) [21]; 5000 Insect Genome Project (i5K) [22]; 1000 Plants Project (1KP) [23]; Oz Mammalian Genomics [24]; Earth BioGenome Project [25]). Readily available conspecific reference genomes for threatened species will likely enable faster uptake of a conservation genomics approach, for example, by avoiding the time and expenditure of sequencing and assembling a high-quality genome de novo. However, in many instances, the next best available resource may be a genome from a closely related species. There has been discussion on the utility of closely related reference genomes for reference-guided genome assembly (i.e., [26,27]). Additionally, there are many research studies to date that have used closely related reference genomes for SNP discovery using reduced-representation and whole genome resequencing (hereafter, resequencing) approaches (e.g., [28–31]).

Birds offer an exceptional opportunity to study the utility of SNP discovery using closely related reference genomes to inform conservation management. In comparison with other vertebrates, bird genomes are relatively small (~0.93–1.3 Gb), compact (i.e., low repetitive elements), and conserved between species [32,33]. Also, the availability of bird reference genomes has increased, due in part to the efforts of individual research groups that produce genomes to answer questions regarding primary production (e.g., chicken, *Gallus gallus* [34]; the turkey, *Meleagris gallopavo* [35]), evolution (e.g., zebra finch, *Taeniopygia guttata*, [36]; Galapagos cormorant, *Phalacrocorax harrisi* [37]), and conservation (e.g., ‘amakihī/Hawaiian honeycreeper, *Hemignathus virens* [38]; ‘alalā/Hawaiian crow, *Corvus hawaiiensis* [39]; kākāpō, *Strigops habroptilus* [40]; kākī/black stilt, *Himantopus novaezelandiae*, this study). A substantial increase in the number of reference genomes available for birds can also be attributed to the efforts of B10K [21,41], the international consortium whose goal is to produce a genome for every known species of bird. To date, B10K has published 38 de novo bird reference genomes [21]. These genomes, along with others that were available at the time of publication, make genomic resources available for at least one individual in almost every order of class Aves [42]. The next phase of B10K will include genomes representing one species from every bird family ( $n = 240$ , [42]), increasing the availability of conspecific or closely related reference genomes for conservation research.

Here, we explore the utility of closely related reference genomes for SNP discovery using a critically endangered wading bird, the kakī, as proof of concept. Once found on the North and South Islands of New Zealand, kakī experienced significant population decline throughout the 20th century due to habitat loss and degradation, and the introduction of mammalian predators. Today, there are approximately 132 kakī remaining (New Zealand Department of Conservation, *unpublished data*) and the population is contingent upon intensive management [43,44], including a captive breeding and rearing programme that uses genetic-based estimates of relatedness to pair distantly related individuals in captivity [45]. Beyond kakī, many programmes for threatened species incorporate neutral genetic measures (e.g., nucleotide diversity, individual heterozygosity or inbreeding, and relatedness) into management plans to minimise inbreeding [46] and loss of diversity [47,48] to reduce extinction risk [49,50].

To demonstrate that closely related reference genomes can yield sufficient SNPs to estimate diversity measures in threatened species, we map kakī genotyping-by-sequencing (GBS) and resequencing reads to genomes from members across the order Charadriiformes, including a conspecific reference genome (kakī, family: Recurvirostridae, *H. novaezelandiae*), and members of the same genus (pied stilt, family: Recurvirostridae, *H. himantopus*), family (pied avocet, family: Recurvirostridae, *Recurvirostra avosetta*), and order (killdeer, family: Charadriidae, *Charadrius vociferus*) (Figure 1). Members from this comparison represent a wide evolutionary time scale: estimates based on traditional single-locus phylogenetic approaches suggest Charadriidae and Recurvirostridae diverged approximately 69 million years ago, avocets (genus: *Recurvirostra*) and stilts (genus: *Himantopus*) diverged approximately 36.9 million years ago, and kakī and pied stilt diverged approximately 1 million years ago [51,52] (but see [53]) (Figure 1). SNPs discovered from these reference-guided assemblies were then compared using estimates of diversity relevant to the conservation management of threatened species, including nucleotide diversity, individual heterozygosity, and relatedness.



**Figure 1.** Evolutionary relationships between species with reference genomes used in this proof of concept. The evolutionary tree indicates topology between taxa within the order Charadriiformes. Evolutionary tree is not to scale.

## 2. Materials and Methods

### 2.1. Tissue Sampling and DNA Extractions

Kākī blood samples were collected during routine health checks by the New Zealand Department of Conservation (DOC) at the captive breeding facilities in Twizel (DOC) and Christchurch (Isaac Conservation and Wildlife Trust), New Zealand, by approval of the DOC Animal Ethics Committee (AEC #283). These samples were stored in 95% molecular grade ethanol at  $-80^{\circ}\text{C}$  at the University of Canterbury. Pied stilt blood samples were collected from one female and one male during routine health checks at Adelaide Zoo, with samples provided under the Royal Zoological Society of South Australia Specimen Licence Agreement (Import Permit: 2016061954). Pied stilt samples were stored in EDTA at  $-20^{\circ}\text{C}$  at the University of Otago. The pied avocet blood sample was collected from a single individual from Hamburger Hallig, Germany, under a permit from the Ministry of Energy, Agriculture, the Environment, Nature and Digitization of the federal state of Schleswig-Holstein, Germany (Permit: V312-7224.121-37 [42-3/13]). Pied avocet samples were stored on filter paper at  $-20^{\circ}\text{C}$  at the University of Kiel.

Genomic DNA for kākī and pied stilt reference genomes was extracted at the University of Otago using a Thermo Scientific™ MagJET™ Genomic DNA Kit (Waltham, USA) following manufacturer specifications. DNA was isolated for the pied avocet sample at the University of Kiel Institute for Clinical Molecular Biology (hereafter, IKMB) by adding 400  $\mu\text{L}$  of phosphorus buffered saline solution (PBS) to dried blood and using the Qiagen® QIAmp® DNA Blood Mini QIAcube® Kit (Hilden, Germany) following the manufacturer specifications. Genomic DNA for the kākī genotyping-by-sequencing (GBS) and resequencing approaches was extracted at the University of Canterbury using a lithium chloride chloroform extraction method (see Supplement 1 for details). Genomic DNA for all extractions were analysed for quality using a NanoDrop™ Spectrophotometer and for quantity using an Invitrogen™ Qubit™ Fluorometer.

### 2.2. Reference Genome Library Preparation and Sequencing

Paired-end libraries for the kākī and pied stilt were prepared at the University of Otago using the Illumina TruSeq® DNA PCR-free protocol according to manufacturer specifications, with genomic DNA fragmented to 350 bp. End repair and adapter ligation for sequence barcoding were carried out and libraries were indexed with unique 6 bp sequences. Sequencing of kākī and pied stilt libraries was completed by New Zealand Genomics Limited (NZGL), where sample libraries were pooled with three additional stilt samples and spread across five lanes of a flow cell for  $2 \times 125$  bp sequencing on an Illumina HiSeq 2500.

Paired-end libraries for the pied avocet were prepared using the TruSeq® DNA Nano Library Prep protocol according to manufacturer specifications, with genomic DNA fragmented to 350 bp. Library preparation and sequencing for the pied avocet was completed at IKMB using one lane of a flow cell on an Illumina HiSeq 4000 for  $2 \times 150$  bp sequencing.

### 2.3. Reference Genome Sequence Processing and Assembly

#### 2.3.1. Kākī and Australian Pied Stilt

Raw kākī and pied stilt sequence reads were evaluated for quality using FastQC v. 0.11.5 [54]. To test for exogenous contamination, the presence and abundance of non-avian reads was estimated by randomly subsampling 5000 reads from each library and searching these reads against the NCBI nucleotide database using BLAST [55].

Illumina adapters used for sequence barcoding were removed using Trimmomatic v. 0.35 [56]. Low quality bases were trimmed using ConDeTri v. 2.3 [57] with default settings. Read deduplication was carried out with ConDeTri, using the first 50 bp of both reads in a pair for comparisons. Raw reads were analysed using SGA-preqc v. 0.9.4 [58] to generate estimates of genome size and heterozygosity.

To determine the level of expected heterozygosity in the genome and assess potential signatures of contamination, paired-end reads were analysed using KmerGenie [59].

Trimmed sequences were assembled with SOAPdenovo2 [60] following initial testing of several assemblers and varying k-mer values. Draft assembly metrics were independently assessed with the assembly metrics script generated for Assemblathon [61]. BUSCO v. 3.0.1 [62,63] was used to determine completeness of the assembly outputs based on expected gene content using an avian ortholog set derived from OrthoDB v. 9 [64] and the chicken as reference. Both assembly metrics and BUSCO scores were used to determine the highest quality assemblies.

Trimmed sequence reads were used to close gaps between scaffolds in the highest quality assemblies for kakī and pied stilt with GapCloser v. 1.12 [60]. Scaffolds shorter than 5 kbp were removed, and genomes were syntenically aligned against the chicken reference genome (version 5.0, GenBank Assembly GCF\_000002315.5) using Chromosome in Satsuma v. 3.1.0 [65] to generate pseudochromosome-level assemblies by aligning the draft assembly scaffolds against the chicken genome, and retaining orthologous regions. Final drafts of kakī and pied stilt genomes are available (see Data Availability section).

### 2.3.2. Pied Avocet

Raw pied avocet sequence reads were evaluated for quality using FastQC v. 0.11.5 [54]. To remove low quality reads, paired-end data was trimmed for Illumina adapter contamination and low quality bases using Skewer v. 0.2.2 [66] with a mean Phred-score of 20, end-trim quality of 30, and a minimum length of 54 bp. Raw reads were analysed with SGA Preqc 0.10.15 [58] and KmerGenie [59] to estimate heterozygosity and potential signatures of contamination. These analyses indicated high expected heterozygosity (0.3%) compared to other birds. To eliminate highly abundant repeats and sequencing errors, a digital normalisation was conducted using Khmer 2.1.1 [67].

Pied avocet trimmed sequences were assembled using Velvet 1.2.10 [68] following initial testing with Meraculous-2D v. 2.2.5.1 [69], which failed to produce a high-quality assembly due to an overabundance of incorrectly merged diplotigs (i.e., contig pairs that share a unique k-mer at both ends [70]). To evaluate the misassemblies, a second assembly was done with Velvet using default parameters. All contigs were aligned against the assembly using LAST [71], with the -uNEAR seeding parameter. Alignments were filtered for trivial self-vs-self perfect alignments, with only single high-scoring pairs per sequence over 99% identical kept. These alignments revealed an unusual number of large and frequent indels (> 3 bp, higher than the default Velvet parameter for allowed gaps in graph bubbles) in extremely similar contigs, and therefore a final Velvet assembly was run with adjusted parameters (-ins\_length 410, -max\_branch\_length 50, -max\_divergence 0.1, -max\_gap\_count 10).

Assembled scaffolds were analysed with GapCloser v. 1.12 [60] to decrease gaps in the assembly. The gap-closed assembly was then aligned against the chicken genome using LAST [71] and the Chromosomer [72] toolkit was used to construct superscaffolds. The final draft of the pied avocet genome is available (see Data Availability section).

### 2.3.3. Killdeer

A killdeer genome was published in the ordinal phase of the B10K project [21]. To improve the assembly, a full de novo approach was used to construct a low-level base-accurate assembly. The data used in the original assembly of killdeer was downloaded from the GigaDB website [73]. This consisted of 12 libraries of Illumina sequence data, including five paired-end libraries with insert sizes ranging from 170 bp to 800 bp and seven mate-pair libraries of insert sizes ranging from 2000 bp to 20,000 bp.

FastQC v. 0.11.5 [54] was used to evaluate the quality of the Illumina data, as well as assess the contamination levels present in the samples. All paired-end libraries consisted of paired 100 bp reads, whereas mate-pair libraries were constructed of paired 50 bp reads. There was no evidence of any significant DNA contamination, but the per-base Phred-scores showed a consistently lower quality early in the reads. Due to the issues observed in the FastQC reports, reads were trimmed using Skewer

v 0.2.2 [66] to a minimum Phred-score of 30 and any read pair where at least one of the mates was trimmed to a length of < 32 bp was discarded.

Trimmed sequences were assembled using AllPaths-LG [74,75] following initial testing of several assemblers and varying k-mer values. The first run was made with the two 170 bp libraries and the complete collection of mate-pair libraries. As part of the AllPaths-LG pipeline, a set of diagnostic data was generated, including estimates of genome size, error rates, and SNP rates. Three of the mate-pair libraries were removed from subsequent analysis after low levels of utilisation were detected due to failed library construction.

The new draft assembly was aligned against the original killdeer reference genome produced by Zhang et al. [21] using the program LAST [71], which identified areas of conflict between the original and new draft killdeer genomes (e.g., short gaps, abundance of small indels, and poor resolution in heterozygous regions in the original genome). A custom set of scripts, ‘SemHelpers’ [76], was written to consolidate the changes detected via the genome-wide alignments into the existing reference genome. The resulting assembly has almost identical metrics when compared to the original assembly [21], given the method used. Post-correction alignments between the final assembly and the original reference genome show identities between 98 and 99%.

Quality of all final draft assemblies was assessed with the Assemblathon metrics script [61] and completeness assessed with BUSCO v. 3.0.1 [62,63] using the avian ortholog set and the chicken as reference. The final draft of the killdeer genome is available (see Data Availability section).

#### 2.4. Genotyping-by-Sequencing

Genotyping-by-sequencing (GBS), a reduced-representation genomics approach, was used to produce genome-wide SNPs for kaki. Briefly, GBS reduces genome complexity by sequencing regions that flank restriction enzyme cut sites [77]. The GBS data presented here were produced following the Elshire et al. [77] method, using 50 ng of genomic DNA with 0.72 ng of total adapters and the restriction enzyme ApeKI.

Because the kaki samples were collected during two different breeding seasons, library preparation and sequencing were completed in two separate batches. The first batch included captive parents and offspring from the 2015/2016 breeding season and other individuals sampled from 2014–2015 that represent diverse lineages based on the kaki pedigree ( $n = 52$ ; pedigree data not shown). This batch was sequenced with paired-end,  $2 \times 100$  bp reads on one lane of an Illumina HiSeq 2500 through NZGL. The second batch consisted of captive parents and offspring from the 2016/2017 breeding season plus one wild individual sampled in 2014 who represented a diverse lineage based on the pedigree ( $n = 47$ ). This batch was sequenced with paired-end,  $2 \times 150$  bp reads on one lane of an Illumina X Ten through CustomScience, Ltd. To assess the impact of batch effects (i.e., library and lane biases [78]), 10 individuals were represented in both batches to ensure similar genetic distance estimates were produced by each duplicated sample independently (see Table S1 for individual sample sequencing details).

FastQC v. 0.11.4 [54] was used to evaluate the quality of the raw Illumina data, as well as assess the contamination levels present in the samples. Paired-end reads were demultiplexed and barcodes were trimmed using Axe [79] with a maximum mismatch of 1. To minimise batch effects [78] and address sequence quality, reads from the 2016/2017 breeding season were trimmed to a maximum length of 100 bp using Skewer [66]. To remove low quality data, reads were filtered to discard short reads (< 32 bp) and reads with mean quality scores less than 30.

In order to be read by downstream pipelines, new single-end barcodes were generated for the ApeKI enzyme using the programme GBSX [80] and appended to the forward-end of reads through a custom Perl script, ‘mux\_barcode’ [81]. For this study, the Tassel 5.0 [82] pipeline was used for SNP discovery and genotyping with GBS data. Due to the double-barcoding scheme of the GBS data generated here, a new class of enzymes was created specifically for Tassel 5 to add the enzyme cut site remnant, together with the reverse barcodes, as recognition sites for these datasets. The Tassel

5.0 GBSv2 pipeline was used with tag database and export plugins specifying a k-mer length of 64, a minimum k-mer length of 20, a minimum Phred-score of 30, and a minimum tag count of 10. Bowtie2 [83] was used to align tags to the each draft reference genome using the *–very-sensitive* presetting. The Tassel 5.0 GBSv2 discovery SNP caller plugin was run with a minimum minor allele frequency (*-mnMAF*) of 0.05 and a minimum locus coverage (*-mnLCov*) of 0.75. VCFtools v. 1.9 [84] was used to filter the dataset to a set of bi-allelic SNPs, with an average minimum SNP depth of 5, and 90% of all SNPs being shared amongst individuals. To minimise statistical bias of linkage disequilibrium, the data set was pruned for linkage disequilibrium using BCFtools v. 1.9 [85] with  $r^2$  set to 0.8 and a window size of 1000 sites. To ensure a more even spread of SNPs throughout the genome, VCFtools v. 1.9 [84] was used to reduce the number of SNPs to 1 SNP within 64 bp, which is the designated size of a GBS tag using Tassel 5.0. VCFs of the filtered data set are available (see Data Availability section).

In order to evaluate whether the same SNPs were likely to be mapped using different reference genomes, a custom script, ‘pancompare’ [86], was used to compare pairs of tags in SAM files that are unique or shared between Tassel 5.0 runs using different reference genomes. This method uses tag pair mapping as a proxy for SNP discovery, under the assumptions that tags all start at the restriction cut site and intersecting pairs of tags are likely to discover the same SNPs using different reference genomes.

## 2.5. Resequencing

In addition to a reduced representation approach, we also resequenced kakī genomes from 36 individuals for SNP discovery and genotyping. These individuals include parents and offspring from the 2015/2016 and 2016/2017 breeding seasons ( $n = 24$ ) and other individuals sampled between 2014–2017 that represent diverse lineages based on the pedigree ( $n = 12$ ). Libraries were prepared at IKMB using a TruSeq® Nano DNA Library Prep kit following the manufacturer’s specifications. Libraries were sequenced across 34 lanes on a HiSeq 4000 at the IKMB.

FastQC v. 0.11.4 [54] was used to initially evaluate the quality of the raw Illumina data, as well as assess the contamination levels present in the samples. Reads were trimmed for the Illumina barcode and for a Phred-score of 20 using Trimmomatic [56]. Reads were mapped to each indexed genome using Bowtie2 [83] with the *–very-sensitive* presetting. Resulting SAM files were converted to BAM files and read coverage was analysed using mpileup with Samtools v. 1.9 [85]. To improve the computational efficiency of mpileup, a custom Perl script ‘split\_bamfile\_tasks.pl’ [87] was created to subdivide BAM files and run them in parallel. SNPs were detected, filtered, and reported using BCFtools v.1.9 [85]. Filtering settings included biallelic SNPs with a minor allele frequency  $>0.05$ , an average mean depth  $>10$ , and a Phred-score  $>20$ . BCFtools was used to filter for a maximum of 10% missing data per site. Resulting SNPs were pruned for linkage disequilibrium using BCFtools with  $r^2$  set to 0.8 and a window size of 1000 sites. To ensure a more even spread of SNPs throughout the genome, VCFtools v. 1.9 [84] was used to reduce the number of SNPs to 1 marker within 150 bp, which is the length of resequencing reads. VCFs of the filtered data set are available (see Data Availability section).

## 2.6. Diversity Estimates

Nucleotide diversity ( $\pi$ ) and individual heterozygosity ( $H_S$ ) were estimated using VCFtools v. 1.9 [84]. Pairwise relatedness ( $R$ ) matrices were produced using KGD [88], a programme that estimates relatedness while taking into account read depth of HTS data. Pairwise  $R$  values were scaled so that self-relatedness of all individuals was equal to 1 using the formula:

$$M_S = D \times M_O \times D$$

where  $M_S$  is the scaled matrix,  $M_O$  is the original matrix, and  $D$  is a diagonal matrix with elements:

$$D = 1/\sqrt{(\text{diag}(M_O))}$$



To compare  $H_S$  estimates generated from different reference genomes using GBS and resequencing data, analysis of variance (ANOVA) and Tukey multiple comparisons of means tests were performed using a linear mixed effects model with lme4 [89] to account for repeated measures (i.e., repeated individuals mapped to all four reference genomes). Mantel tests with 1000 iterations were used to test whether scaled pairwise  $R$  matrices using different reference genomes were significantly similar compared to a null distribution. Correlations were conducted between estimates of  $H_S$  and  $R$  (not including self-relatedness) using different reference genomes using Spearman's rank ( $r_s$ ), which accounts for the inherently non-normal distribution of the  $R$  estimates.

### 3. Results

#### 3.1. Reference Genome Sequencing and Assembly

Library sequencing produced 226–307 million paired-end sequences for each kakī, pied stilt, and avocet sample. Average sequencing depth was  $52\times$  for kakī,  $51\times$  for pied stilt, and  $70\times$  for avocet, based on an expected genome size of 1.2 Gb. Genomes produced were between 1.02–1.22 Gb in total length (Table 1), which is within the expected range for avian genomes [90]. Scaffold N50 sizes ranged from 3.66 to 105.71 Mb. The total number of scaffolds ranged from 67 to 15,167. BUSCO assessment indicated the presence of at least 82.4% of the orthologs from the avian database. Combined, these estimates indicate that the assembled genomes have high genome completeness.

**Table 1.** Genome assembly metrics for the genomes assembled in this study.

Species	Total Assembly Length (Gb)	Total Scaffolds	Scaffold N50 (bp)	Longest Scaffold (bp)	Average Scaffold Length (bp)	Complete Single-Copy BUSCOs (%)
Kakī	1.18	523	105,710,992	238,324,410	2,254,638	91.0
Pied Stilt	1.12	1443	99,457,149	221,521,436	773,955	85.9
Avocet	1.02	67	87,059,367	184,945,080	15,204,176	82.4
Killdeer	1.22	15,167	3,657,525	21,923,840	80,436	92.5

#### 3.2. SNP Discovery and Diversity Estimates—GBS

After demultiplexing and initial read filtering, kakī GBS sequencing resulted in a total of 802.4 million reads for 88 individuals (mean =  $9.1 \pm$  S.D. 4.9 million reads per individual). Five of these individuals were subsequently removed from the study after SNP filtering for having low average sample depths across sites ( $<4\times$  depth using conspecific reference genome). The resulting 82 individuals have an average depth of  $11.71\text{--}18.51\times$ , with average missingness of 2–4% depending on the reference genome used (Table 2).

**Table 2.** Mapping statistics, single nucleotide polymorphisms (SNPs) discovered, SNP descriptive statistics, and average diversity statistics from genotyping-by-sequencing (GBS) reads mapped to different reference genomes.  $\pi$ : nucleotide diversity,  $H_S$ : individual heterozygosity,  $R$ : pairwise relatedness ( $\pm$  S.D. for each measure).

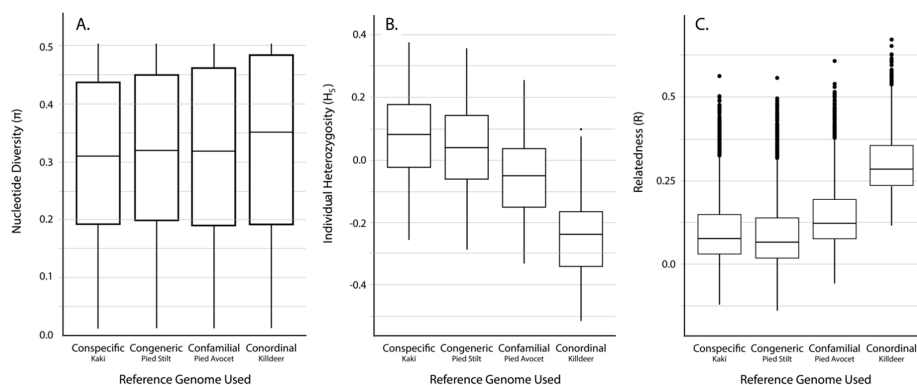
Reference Genome	No. of Mapped Tag Pairs	% Tags Shared with Kakī Mapping	No. Unfiltered SNPs	No. Filtered SNPs	Average Missingness	Average Depth	Average $\pi$	Average $H_S$	Average $R$
Kakī	392,652	100	634,695	19,396	$0.04 \pm 0.04$	$13.73 \pm 6.53$	$0.31 \pm 0.14$	$0.07 \pm 0.15$	$0.11 \pm 0.12$
Pied Stilt	372,906	91.04	604,573	18,625	$0.04 \pm 0.04$	$11.71 \pm 5.52$	$0.32 \pm 0.14$	$0.03 \pm 0.15$	$0.10 \pm 0.12$
Avocet	316,978	83.10	481,532	18,398	$0.03 \pm 0.04$	$13.90 \pm 6.58$	$0.31 \pm 0.15$	$-0.06 \pm 0.14$	$0.15 \pm 0.11$
Killdeer	151,546	72.42	242,493	10,440	$0.02 \pm 0.03$	$18.51 \pm 8.77$	$0.33 \pm 0.15$	$-0.25 \pm 0.14$	$0.30 \pm 0.09$

The number of GBS tag pairs mapped to each reference genome was greatest using a conspecific reference genome, with fewer tag pairs mapped the more phylogenetically distant the reference genome became (Table 2). Results from our analysis with 'pancompare' [86] indicate that more tags from the congeneric mapping were shared with those mapped to a conspecific reference genome



(91.04%) than more distantly related genomes (confamilial = 83.10% and conordinal = 72.42%; Table 2). Tag pairs always start at the GBS restriction enzyme cut site, making direct comparisons of tags mapped across different genomes possible. Because more mapped tags were shared between closely related genomes, these results suggest that SNPs discovered with the congeneric reference genome are more likely the same as those discovered with the conspecific reference genome than those discovered with the confamilial or conordinal references.

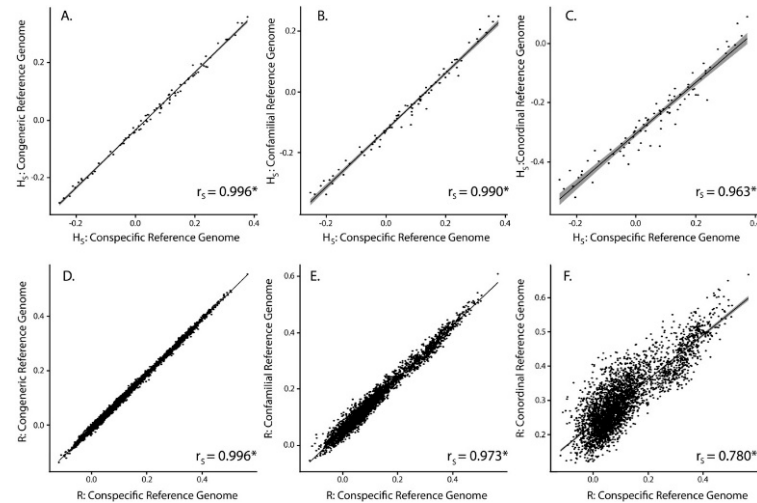
The number of unfiltered and filtered SNPs discovered was greatest when using a conspecific reference genome, with fewer SNPs discovered the more phylogenetically distant the reference genome became (Table 2). Despite the differences in number of SNPs discovered with each reference genome, average nucleotide diversity ( $\pi$ ) was similar across datasets (average  $\pi$  = 0.31–0.33, Table 2, Figure 2A).



**Figure 2.** Distribution of different diversity estimates using SNPs discovered with GBS reads mapped against different reference genomes. (A) Nucleotide diversity ( $\pi$ ), (B) individual heterozygosity ( $H_s$ ), and (C) pairwise relatedness ( $R$ ) not including self-relatedness.

Average individual heterozygosity ( $H_s$ ) estimates differed depending on the reference genome used (Table 2, Figure 2B). Results show that using different reference genomes produced significantly different levels of  $H_s$  from one another (ANOVA,  $p < 0.001$ ; Tukey Contrasts,  $p < 0.001$ ). Using a congeneric reference genome resulted in  $H_s$  estimates that are on average 3.4% less than using a conspecific reference genome, with a confamilial being 12.9% less, and a conordinal being 31.6% less. Despite significant differences in  $H_s$  depending on the reference genome used, estimates of  $H_s$  using different reference genomes were significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.996$ ) being higher than the conspecific and confamilial approaches ( $r_s = 0.990$ ) and conspecific and conordinal approaches ( $r_s = 0.963$ ; Figure 3A–C).

The range of scaled average pairwise estimates of relatedness ( $R$ ) shows a bimodal distribution, which reflects highly related individuals (siblings and parent-offspring relationships) along with more distantly related individuals that are captured in the study design. The range of scaled  $R$  values appeared different depending on the reference genome used, with average pairwise  $R$  in the conspecific and congeneric analyses being less than the confamilial and conordinal analyses (Table 2). Despite this pattern, the relationships between  $R$  using a conspecific reference genome and all other genomes were not significantly different (Mantel test,  $p < 0.001$ ). Estimates of pairwise  $R$  (not including self-relatedness) using different reference genomes were significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.996$ ) being higher than the conspecific and confamilial approaches ( $r_s = 0.973$ ) and the conspecific and conordinal approaches ( $r_s = 0.780$ ; Figure 3D–F).



**Figure 3.** Scatterplots showing individual point estimates of  $H_S$  (A–C) and pairwise  $R$  estimates (D–F) using GBS reads mapped to different reference genomes. Self-relatedness estimates were not used in this analysis. Trend line in black, with 95% confidence intervals surrounding the trend line in gray. Spearman’s correlation coefficient ( $r_s$ ) provided in the lower right corner of each scatterplot. \* indicates significance  $p < 0.001$ .

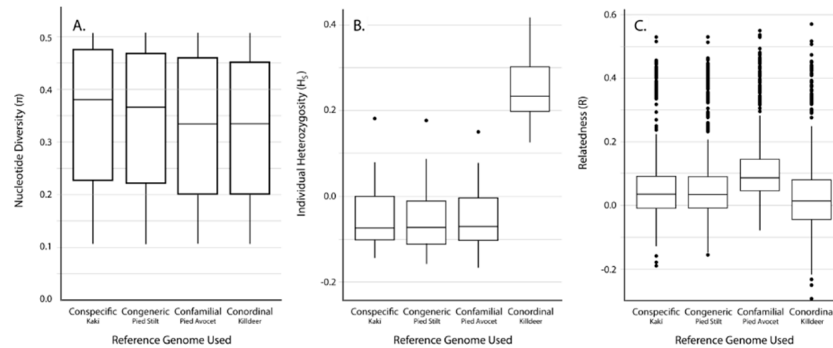
### 3.3. SNP Discovery and Diversity Estimates—Resequencing

After demultiplexing and initial read filtering, the kakī resequencing resulted in a total of 4.8 billion reads for 36 individuals (mean =  $135.8 \pm 54.1$  million reads per individual). After SNP filtering, these 36 individuals have an average depth of 13.95–17.44 $\times$  with average missingness of 0.2% across all reference genomes used (Table 3).

**Table 3.** Alignment rates, single nucleotide polymorphisms (SNPs) discovered, SNP descriptive statistics, and average diversity statistics from resequencing reads mapped to different reference genomes.  $\pi$ : nucleotide diversity,  $H_S$ : individual heterozygosity,  $R$ : pairwise relatedness. ( $\pm$  S.D. for each measure).

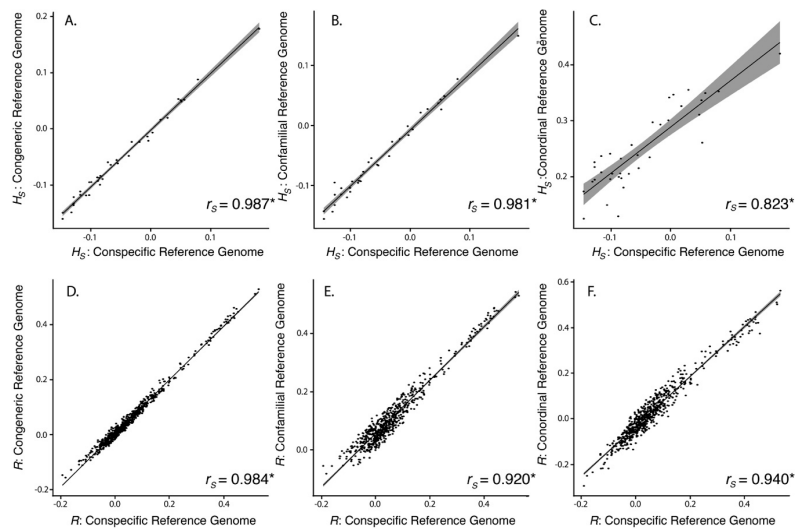
Reference Genome	Average Alignment Rate (%)	No. Unfiltered SNPs	No. Filtered SNPs	Average Missingness	Average Depth	Average $\pi$	Average $H_S$	Average $R$
Kakī	$94.6 \pm 0.50$	4,246,100	91,854	$0.002 \pm 0.005$	$17.44 \pm 6.79$	$0.35 \pm 0.13$	$-0.05 \pm 0.08$	$0.06 \pm 0.11$
Pied Stilt	$88.1 \pm 0.96$	8,438,866	89,419	$0.002 \pm 0.005$	$14.99 \pm 6.06$	$0.34 \pm 0.13$	$-0.05 \pm 0.08$	$0.06 \pm 0.11$
Avocet	$78.5 \pm 0.46$	24,333,620	143,343	$0.002 \pm 0.004$	$16.02 \pm 6.43$	$0.33 \pm 0.14$	$-0.05 \pm 0.07$	$0.11 \pm 0.11$
Killdeer	$64.8 \pm 4.89$	62,888,931	89,145	$0.002 \pm 0.004$	$13.95 \pm 5.54$	$0.32 \pm 0.13$	$0.25 \pm 0.07$	$0.03 \pm 0.13$

Average read alignment rates using Bowtie2 were highest when using a conspecific reference genome (94.6%), with fewer reads aligning with congenic (88.1%), confamilial (78.5%), and conordinal reference genomes (64.8%, Table 3). In contrast to GBS, the number of unfiltered SNPs increased with phylogenetic distance of the reference genome, which is expected given resequencing SNPs are called by differences between reads and the reference. The number of SNPs discovered post filtering did not correspond with phylogenetic distance of the reference used, with the fewest filtered SNPs being discovered with the conordinal reference genome (89,145) and the most being discovered with the confamilial reference genome (143,343, Table 3). Similar to the GBS dataset, average  $\pi$  was similar across datasets generated using different reference genomes (average  $\pi = 0.32$ – $0.35$ , Table 3, Figure 4A).



**Figure 4.** Distribution of different diversity estimates using SNPs discovered with resequencing reads mapped against different reference genomes. (A) Nucleotide diversity ( $\pi$ ), (B) individual heterozygosity ( $H_S$ ), and (C) pairwise relatedness ( $R$ ). Self-relatedness estimates were not used in this analysis.

Results show that using a conordinal reference genome produced significantly higher levels of  $H_S$  than the conspecific, congeneric, or confamilial approaches (ANOVA,  $p < 0.001$ ; Tukey contrasts,  $p < 0.001$ ; Table 2, Figure 4B). Using a congeneric reference genome resulted in  $H_S$  estimates that are on average 0.40% less than using a conspecific reference genome, with a confamilial being 0.31% less, and a conordinal being 29.9% greater. Despite significant differences in  $H_S$  depending on the reference genome used,  $H_S$  using different reference genomes is significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.987$ ) being higher than congeneric and confamilial approaches ( $r_s = 0.981$ ) and congeneric and conordinal approaches ( $r_s = 0.823$ ; Figure 5A–C).



**Figure 5.** Scatterplots showing individual point estimates of  $H_S$  (A–C) and pairwise  $R$  estimates (D–F) using resequencing reads mapped to different reference genomes. Self-relatedness estimates were not used in this analysis. Trend line in black, with 95% confidence intervals surrounding the trend line in gray. Spearman's correlation coefficient ( $r_s$ ) provided in the lower right corner of each scatterplot. \* indicates significance  $p < 0.001$ .

Similar to the GBS approach, the range of scaled average pairwise estimates of relatedness ( $R$ ) based on resequencing also shows a bimodal distribution, which reflects the relationships of individuals captured in the study design. Average scaled pairwise estimates of  $R$  were similar across all reference genomes used (Table 2, Figure 4C). The relationship between  $R$  using a conspecific reference genome and all other genomes were not statistically different compared to the null distribution (Mantel test,  $p < 0.001$ ). Scaled pairwise  $R$  (not including self-relatedness) using different reference genomes is significantly correlated (Spearman's correlation,  $p < 0.001$ ), with correlation coefficients between the conspecific and congeneric approaches ( $r_s = 0.984$ ) being higher than conspecific and confamilial approaches ( $r_s = 0.920$ ) and conspecific and conordinal approaches ( $r_s = 0.940$ ; Figure 5D–F).

#### 4. Discussion

For species of conservation concern, limited conspecific genomic resources often impede inclusion of genomic data in conservation management strategies. Our proof of concept demonstrates that SNPs discovered using congeneric, confamilial, and even conordinal approaches yield diversity estimates that significantly correlate with estimates derived from SNPs discovered using a conspecific approach. Prior to this study, there was only one genome publicly available for the order Charadriiformes (i.e., the killdeer [21]). This study provides three additional high-quality de novo genome assemblies, all of which have practical applications for conservation.

The number of GBS tag pairs that aligned to each reference genome decreased the more phylogenetically distant the reference genome became. Because Tassel 5.0 calls SNPs based on differences among tag pairs [82]—as opposed to differences between tag pairs and the reference genome—the number of unfiltered SNPs discovered also decreased the more phylogenetically distant the reference genome became. The same pattern was observed for the number of filtered SNPs. The ‘pancompare’ analysis of GBS tag data suggests that SNP discovery using the conspecific and congeneric reference genomes are more likely to yield the same markers compared to SNPs discovered using the confamilial or conordinal reference genomes.

The number of resequencing reads that aligned to each reference genome also decreased the more phylogenetically distant the reference genome became. Unlike GBS, the number of unfiltered SNPs increased with phylogenetic distance. This is to be expected because the resequencing discovery pipeline calls SNPs based on differences between reads and the reference genome [85]. The number of SNPs discovered post-filtering was unexpected, however, as a similar number of SNPs were found in all but the confamilial reference approach, which resulted in  $\sim 1.5 \times$  more SNPs than other reference-guided approaches. While the pied avocet genome shows signs of high completeness, complexities in the genome assembly due to high heterozygosity [69,91] may have resulted in less complete regions leading to higher false discovery rates [41].

Using GBS and resequencing data, the average and range of  $\pi$  estimates did not differ greatly based on reference genome used. Larger differences between reference genomes used were observed when estimating  $H_S$ . Using GBS data, mean estimates of  $H_S$  decreased significantly the more distant the reference genomes became, with the use of a conordinal reference genome producing a marked decrease in  $H_S$  estimates compared to the use of a conspecific reference. This decrease in  $H_S$  corresponds to an increase in  $R$ , although not significantly so. These combined results are consistent with expectations because SNPs called by Tassel 5.0 are based on identifying mapped tag pairs [82]; the more phylogenetically distant a reference genome is, the more conserved a region has to be to successfully map a pair of tags. Therefore, with GBS we expect  $H_S$  to be lower and  $R$  to be higher the more phylogenetically distant the reference used is, given that variants at these conserved regions are less frequent.

Using resequencing data, conspecific, congeneric, and confamilial approaches produced  $H_S$  that were not significantly different from one another, with the only significant difference seen with the conordinal approach, which resulted in a significant increase in  $H_S$  compared to other reference genomes. Unlike GBS tags, there is not an immediate explanation for this pattern. However, it may

be attributed to the fact that resequencing reads, which are longer and are more representative of the whole genome, can be mapped to more divergent regions than GBS tags.

While the range of  $H_S$  and pairwise  $R$  values may be different depending on the reference genome used, all estimates produced using different reference genomes correlate significantly with one another. Our results suggest that using a more closely related reference genome (e.g., congeneric) over a more distant reference genome (e.g., conordinal) will yield SNPs that have higher correlation coefficients with estimates generated using a conspecific, and therefore, are likely to result in similar conservation recommendations. Ongoing work incorporating genomic based estimates of relatedness into software that informs captive pairing recommendations (e.g., PMx [92]) will indicate whether more distantly related reference genomes indeed produce statistically similar pairing recommendations, as our correlation results suggest. In the meantime, we anticipate even small changes in  $H_S$  and pairwise  $R$  estimates will not greatly affect conservation recommendations, as diversity estimates are often used in relative terms. For example, pairing recommendations for intensively managed populations that lack reliable pedigrees are routinely informed by genetic- or genomic-based pairwise estimates of relatedness (e.g., [45,93–95]). In practice, pairing recommendations are made based on the relative ranking of these estimates and not the absolute values. Similarly, when investigating heterozygosity-fitness correlations (e.g., [96]), relative rankings of  $H_S$  among individuals are more informative than absolute values.

Still, there may be some instances where absolute diversity values may be of interest (e.g., parentage assignment, or management of individuals that exhibit  $H_S$  below a cutoff score [97]). SNPs derived using the conordinal reference genome provide markedly different ranges of  $H_S$  and pairwise  $R$  estimates and often the lowest correlation coefficients compared to SNPs derived from the conspecific reference genome. For birds, we recommend a confamilial reference genome as the most distant reference genome conservation researchers consider using for diversity estimates. However, this approach should be evaluated for use in other questions, such as the characterisation of adaptive variation [4,14].

The number of de novo bird genomes available to be used as reference is due to increase, especially as the next phase of B10K seeks to publish representative genomes for every recognised family of birds [42]. However, we recommend evaluating the quality of publicly available genomes prior to use, as lower quality genomes may produce lower SNP yield due to fewer alignable regions, or greater false discovery rate where there are assembly errors [98]. Here, we re-assembled the available killdeer reference genome for two reasons. First, the raw data available from the European Bioinformatics Institute European Nucleotide Archive (EBI ENA) showed poor sequencing quality and mapping of this raw data to the existing reference suggested inconsistencies where poor quality reads were more abundant. Second, mapping of the long-insert mate-pair data from the project showed little to no support for many of the scaffolding connections present in the published genome. Due to these factors, we reassembled the genome using much more stringent data curation and more cautious scaffolding. Given this, when using a genome “off the shelf”, we recommend careful assessment of the original genome publication, keeping in mind that genomes assembled from multiple libraries or data types, with greater depth of sequencing coverage, and a more complete and contiguous assembly, will be of higher quality [41]. When genomes with similar phylogenetic relationships are available, comparisons of synteny [65] and completeness [41] against the most closely related model genome may help identify which genome is most appropriate to use. Ultimately, the best way to assess existing genomic resources is to download the raw reads and evaluate them using tools such as FastQC [54] and SGA pre-QC [58], as we have done with the killdeer genome. Raw read quality may have the largest impact on final assembly quality, and initial quality checks will allow identification of any potential anomalies or limitations of the raw data that may have presented challenges to assembly, such as high heterozygosity [69,91,99]. If the raw data is of high quality, but there are inconsistencies between original reported statistics and those derived from raw reads, it may be worth investing in re-assembly to produce a genome of higher quality with greater confidence.

Indeed, re-assembly remains a more cost-effective option than starting a genome sequencing project from scratch. By our current (2018) estimates based on single libraries with paired-end reads, the use of a closely related high quality readily-available reference genome may save a conservation genomic project a minimum of EUR 6500 in library preparations, sequencing, computational power, and assembly time (Table S2, although prices subject to rapid change given new sequencing technologies). Among the 383 species in the order Charadriiformes, 51 are threatened with extinction [100]. The families Laridae (gulls, terns, and skimmers) and Scolopacidae (sandpipers) contain particularly high numbers of threatened species (14 and 13, respectively). Along with the genomes produced in this paper, there are now genomes available for four additional families within Charadriiformes (i.e., Alcidae [90], Charadriidae ([21], here), Recurvirostridae (here), and Scolopacidae [101]). Genome sequencing and assembly of one member of the Laridae family could benefit all 14 threatened species within this family, and combined with the existing genomes available as reference within Scolopacidae, could save conservation groups up to EUR 169,000 in sequencing and assembly costs. Using existing genomic resources will not only reduce these costs, but also the time needed to produce a high-quality reference genome, thereby allowing for a faster uptake of conservation genomics approaches to produce robust information for conservation management.

## 5. Conclusions

Many threatened species management programmes rely on measures of diversity, including nucleotide diversity, heterozygosity, and relatedness, in guiding management decisions [93,102]. While these measures have historically been calculated using small numbers of genetic markers, genomic markers offer the opportunity for increased resolution [1,6,103] and hence improved decision-making. Here, we have demonstrated that in the absence of a conspecific reference genome to map genomic sequence reads to, the availability of high-quality reference genome for a closely related species can provide highly correlated estimates for nucleotide diversity, individual heterozygosity, and relatedness. We anticipate the use of readily available reference genomes may provide resource-constrained conservation projects a way to minimise these costs and make a faster transition to using genomic data to improve conservation outcomes for threatened species.

**Data Availability:** The pied stilt Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession RSEF00000000. The version described in this paper is version RSEF01000000. The pied avocet genome raw reads have been deposited in Genbank under project number PRJNA508299. The reassembled killdeer genome is available at <http://www.uconsert.org/data/>. Kākī are taonga (treasured) to Māori (the indigenous people of Aotearoa New Zealand), and as such the genomes obtained from taonga species are taonga in their own right. Therefore, the genome for kākī and all VCFs for GBS and resequencing will be made available on recommendation of the iwi (tribes) that affiliate as kaitiaki (guardians) for kākī. A local genome browser is available to view the kākī genome and all VCFs presented here at <http://www.uconsert.org/data/>, along with details on how to request access.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2073-4425/10/1/9/s1>, Table S1: Samples used in Genotyping-by-Sequencing and resequencing analyses; Table S2: Cost associated with genome sequencing and alignment; lithium chloride extraction protocol.

**Author Contributions:** All authors contributed to research conceptualization. Genomes presented here were prepared and assembled by N.J.F., R.M. and M.H., G.B.S. and resequencing data was prepared and mapped by S.J.G. with guidance from R.M. Marker discovery and subsequent diversity analyses were performed by S.J.G. Original draft preparation and writing completed by S.J.G. and N.J.F., with review and editing by all authors.

**Funding:** This research was funded by the Brian Mason Scientific & Technical Trust (SJT, TES), the Mohua Charitable Trust (TES), New Zealand Ministry of Business, Innovation and Employment Endeavour Fund (TES, AWS), and the Royal Society of New Zealand Rutherford Discovery Fellowship (MK).

**Acknowledgments:** We are grateful for the support of Te Rūnanga o Ngāi Tahu, Te Rūnanga o Arowhenua, Te Rūnanga o Waihao and Te Rūnanga o Moeraki. We thank Elshire Group Ltd. for their support in generating GBS data. We thank Sabine Müller of the Research and Technology Centre West Coast, Kiel University, for kindly providing access to the pied avocet blood sample, and the Royal Zoological Society of South Australia (Adelaide Zoo) for providing the pied stilt blood sample. Finally, we thank all members of the Kākī Recovery Programme, especially Simone Cleland and Cody Thyne, and the Isaac Conservation and Wildlife Trust, especially Anne

Richardson, for their assistance with sample collection and their ongoing support. We also thank three anonymous reviewers for their helpful comments on the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Allendorf, F.W.; Hohenlohe, P.A.; Luikart, G. Genomics and the future of conservation genetics. *Nat. Rev. Genet.* **2010**, *11*, 697. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Harrison, K.A.; Pavlova, A.; Telonis-Scott, M.; Sunnucks, P. Using genomics to characterize evolutionary potential for conservation of wild populations. *Evol. Appl.* **2014**, *7*, 1008–1025. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Kohn, M.H.; Murphy, W.J.; Ostrander, E.A.; Wayne, R.K. Genomics and conservation genetics. *Trends Ecol. Evol.* **2006**, *21*, 629–637. [\[CrossRef\]](#)
4. Mable, B.K. Conservation of adaptive potential and functional diversity: integrating old and new approaches. *Conserv. Genet.* **2018**, 1–12. [\[CrossRef\]](#)
5. Luikart, G.; England, P.R.; Tallmon, D.; Jordan, S.; Taberlet, P. The power and promise of population genomics: From genotyping to genome typing. *Nat. Rev. Genet.* **2003**, *4*, 981. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Galla, S.J.; Buckley, T.R.; Elshire, R.; Hale, M.L.; Knapp, M.; McCallum, J.; Moraga, R.; Santure, A.W.; Wilcox, P.; Steeves, T.E. Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances. *Mol. Ecol.* **2016**, *25*, 5267–5281. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Shafer, A.B.; Wolf, J.B.; Alves, P.C.; Bergström, L.; Bruford, M.W.; Brännström, I.; Colling, G.; Dalén, L.; De Meester, L.; Ekblom, R.; et al. Genomics and the challenging translation into conservation practice. *Trends Ecol. Evol.* **2015**, *30*, 78–87. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Knight, A.T.; Cowling, R.M.; Rouget, M.; Balmford, A.; Lombard, A.T.; Campbell, B.M. Knowing but not doing: Selecting priority conservation areas and the research-implementation gap. *Conserv. Biol.* **2008**, *22*, 610–617. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Taylor, H.R.; Dussex, N.; van Heezik, Y.J.G.E. Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Glob. Ecol. Conserv.* **2017**, *10*, 231–242. [\[CrossRef\]](#)
10. McCormack, J.E.; Hird, S.M.; Zellmer, A.J.; Carstens, B.C.; Brumfield, R.T. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenetics Evol.* **2013**, *66*, 526–538. [\[CrossRef\]](#)
11. Hayden, E.C. The \$1,000 genome. *Nature* **2014**, *507*, 294–295. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Muir, P.; Li, S.; Lou, S.; Wang, D.; Spakowicz, D.J.; Salichos, L.; Zhang, J.; Weinstock, G.M.; Isaacs, F.; Rozowsky, J.; et al. The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biol.* **2016**, *17*, 53. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Narum, S.R.; Buerkle, C.A.; Davey, J.W.; Miller, M.R.; Hohenlohe, P.A. Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* **2013**, *22*, 2841–2847. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Andrews, K.R.; Good, J.M.; Miller, M.R.; Luikart, G.; Hohenlohe, P.A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **2016**, *17*, 81. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Davey, J.W.; Hohenlohe, P.A.; Etter, P.D.; Boone, J.Q.; Catchen, J.M.; Blaxter, M.L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **2011**, *12*, 499. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Ilut, D.C.; Nydam, M.L.; Hare, M.P.J.B.R.I. Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering. *Biomed Res. Int.* **2014**, *2014*, 675158. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Oyler-McCance, S.J.; Oh, K.P.; Langin, K.M.; Aldridge, C.L. A field ornithologist's guide to genomics: Practical considerations for ecology and conservation. *Auk* **2016**, *133*, 626–648. [\[CrossRef\]](#)
18. Waldron, A.; Mooers, A.O.; Miller, D.C.; Nibbelink, N.; Redding, D.; Kuhn, T.S.; Roberts, J.T.; Gittleman, J.L. Targeting global conservation funding to limit immediate biodiversity declines. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 12144–12148. [\[CrossRef\]](#)
19. Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **2014**, *29*, 51–63. [\[CrossRef\]](#)
20. Genome 10K Community of Scientists. Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J. Hered.* **2009**, *100*, 659–674. [\[CrossRef\]](#)



21. Zhang, G.; Li, B.; Li, C.; Gilbert, M.T.P.; Jarvis, E.D.; Wang, J. Comparative genomic data of the Avian Phylogenomics Project. *GigaScience* **2014**, *3*, 26. [CrossRef] [PubMed]
22. Robinson, G.E.; Hackett, K.J.; Purcell-Miramontes, M.; Brown, S.J.; Evans, J.D.; Goldsmith, M.R.; Lawson, D.; Okamuro, J.; Robertson, H.M.; Schneider, D.J. Creating a buzz about insect genomes. *Science* **2011**, *331*, 1386. [CrossRef] [PubMed]
23. Matasci, N.; Hung, L.-H.; Yan, Z.; Carpenter, E.J.; Wickett, N.J.; Mirarab, S.; Nguyen, N.; Warnow, T.; Ayyampalayam, S.; Barker, M.; et al. Data access for the 1000 Plants (1KP) project. *GigaScience* **2014**, *3*, 17. [CrossRef] [PubMed]
24. Duchêne, D.A.; Bragg, J.G.; Duchêne, S.; Neaves, L.E.; Potter, S.; Moritz, C.; Johnson, R.N.; Ho, S.Y.W.; Eldridge, M.D.B. Analysis of phylogenomic tree space resolves relationships among marsupial Families. *Syst. Biol.* **2017**, *67*, 400–412. [CrossRef] [PubMed]
25. Lewin, H.A.; Robinson, G.E.; Kress, W.J.; Baker, W.J.; Coddington, J.; Crandall, K.A.; Durbin, R.; Edwards, S.V.; Forest, F.; Gilbert, M.T.P.; et al. Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4325–4333. [CrossRef] [PubMed]
26. Card, D.C.; Schield, D.R.; Reyes-Velasco, J.; Fujita, M.K.; Andrew, A.L.; Oyler-McCance, S.J.; Fike, J.A.; Tomback, D.F.; Ruggiero, R.P.; Castoe, T.A. Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. *PLoS ONE* **2014**, *9*, e106649. [CrossRef]
27. Lischer, H.E.; Shimizu, K.K. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinform.* **2017**, *18*, 474. [CrossRef]
28. Der Sarkissian, C.; Ermini, L.; Schubert, M.; Yang, M.A.; Librado, P.; Fumagalli, M.; Jónsson, H.; Bar-Gal, G.K.; Albrechtsen, A.; Vieira, F.G.; et al. Evolutionary genomics and conservation of the endangered Przewalski's horse. *Curr. Biol.* **2015**, *25*, 2577–2583. [CrossRef]
29. Ng, E.Y.; Garg, K.M.; Low, G.W.; Chattopadhyay, B.; Oh, R.R.; Lee, J.G.; Rheindt, F.E. Conservation genomics identifies impact of trade in a threatened songbird. *Biol. Conserv.* **2017**, *214*, 101–108. [CrossRef]
30. Nuijten, R.J.; Bosse, M.; Crooijmans, R.P.; Madsen, O.; Schaftenaar, W.; Ryder, O.A.; Groenen, M.A.; Megens, H.-J. The use of genomics in conservation management of the endangered visayan warty Pig (*Sus cebifrons*). *Int. J. Genom.* **2016**, *2016*, 5613862. [CrossRef]
31. Westbury, M.V.; Hartmann, S.; Barlow, A.; Wiesel, I.; Leo, V.; Welch, R.; Parker, D.M.; Sicks, F.; Ludwig, A.; Dalén, L.; et al. Extended and Continuous Decline in Effective Population Size Results in Low Genomic Diversity in the World's Rarest Hyena Species, the Brown Hyena. *Mol. Biol. Evol.* **2018**, *35*, 1225–1237. [CrossRef] [PubMed]
32. Organ, C.L.; Shedlock, A.M.; Meade, A.; Pagel, M.; Edwards, S.V. Origin of avian genome size and structure in non-avian dinosaurs. *Nature* **2007**, *446*, 180. [CrossRef] [PubMed]
33. Zhang, G.; Li, C.; Li, Q.; Li, B.; Larkin, D.M.; Lee, C.; Storz, J.F.; Antunes, A.; Greenwold, M.J.; Meredith, R.W.; et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **2014**, *346*, 1311–1320. [CrossRef] [PubMed]
34. Consortium, I.C.G.S. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **2004**, *432*, 695.
35. Dalloul, R.A.; Long, J.A.; Zimin, A.V.; Aslam, L.; Beal, K.; Blomberg, L.A.; Bouffard, P.; Burt, D.W.; Crasta, O.; Crooijmans, R.P.; et al. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. *PLoS Biol.* **2010**, *8*, e1000475. [CrossRef] [PubMed]
36. Warren, W.C.; Clayton, D.F.; Ellegren, H.; Arnold, A.P.; Hillier, L.W.; Küstner, A.; Searle, S.; White, S.; Vilella, A.J.; Fairley, S.; et al. The genome of a songbird. *Nature* **2010**, *464*, 757. [CrossRef]
37. Burga, A.; Wang, W.; Ben-David, E.; Wolf, P.C.; Ramey, A.M.; Verdugo, C.; Lyons, K.; Parker, P.G.; Kruglyak, L. A genetic signature of the evolution of loss of flight in the Galapagos cormorant. *Science* **2017**, *356*, eaal3345. [CrossRef] [PubMed]
38. Callicrate, T.; Dikow, R.; Thomas, J.W.; Mullikin, J.C.; Jarvis, E.D.; Fleischer, R.C. Genomic resources for the endangered Hawaiian honeycreepers. *BMC Genom.* **2014**, *15*, 1098. [CrossRef]
39. Sutton, J.; Helmkampf, M.; Steiner, C.; Bellinger, M.R.; Korch, J.; Hall, R.; Baybayan, P.; Muehling, J.; Gu, J.; Kingan, S.; et al. A high-quality, long-read de novo genome assembly to aid conservation of Hawaii's last remaining crow species. *bioRxiv* **2018**, *2018*, 349035.
40. Strigops Habroptilus. 2018. Available online: [https://vgp.github.io/genomeark/Strigops\\_habroptilus](https://vgp.github.io/genomeark/Strigops_habroptilus). (accessed on 7 November 2018).



41. Peona, V.; Weissensteiner, M.H.; Suh, A. How complete are “complete” genome assemblies?—An avian perspective. *Mol. Ecol. Resour.* **2018**, *18*, 1188–1195. [CrossRef]
42. Zhang, G. Genomics: Bird sequencing project takes off. *Nature* **2015**, *522*, 34. [CrossRef] [PubMed]
43. Reed, C.E.M. *Management Plan for Captive Black Stilts*; Biodiversity Recovery Unit, Department of Conservation: Wellington, New Zealand, 1998.
44. Sanders, M.D.; Maloney, R.F. Causes of mortality at nests of ground-nesting birds in the Upper Waitaki Basin, South Island, New Zealand: A 5-year video study. *Biol. Conserv.* **2002**, *106*, 225–236. [CrossRef]
45. Hagen, E.N.; Hale, M.L.; Maloney, R.F.; Steeves, T.E. Conservation genetic management of a critically endangered New Zealand endemic bird: Minimizing inbreeding in the Black Stilt *Himantopus novaezelandiae*. *Ibis* **2011**, *153*, 556–561. [CrossRef]
46. Ford, M.J.; Parsons, K.; Ward, E.; Hempelmann, J.; Emmons, C.K.; Bradley Hanson, M.; Balcomb, K.C.; Park, L.K. Inbreeding in an endangered killer whale population. *Anim. Conserv.* **2018**. [CrossRef]
47. Jordan, S.; Giersch, J.J.; Muhlfeld, C.C.; Hotaling, S.; Fanning, L.; Tappenbeck, T.H.; Luikart, G. Loss of genetic diversity and increased subdivision in an endemic alpine stonefly threatened by climate change. *PLoS ONE* **2016**, *11*, e0157386.
48. Pacioni, C.; Hunt, H.; Allentoft, M.E.; Vaughan, T.G.; Wayne, A.F.; Baynes, A.; Haouchar, D.; Dortch, J.; Bunce, M. Genetic diversity loss in a biodiversity hotspot: Ancient DNA quantifies genetic decline and former connectivity in a critically endangered marsupial. *Mol. Ecol.* **2015**, *24*, 5813–5828. [CrossRef]
49. O’Grady, J.J.; Brook, B.W.; Reed, D.H.; Ballou, J.D.; Tonkyn, D.W.; Frankham, R. Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. *Biol. Conserv.* **2006**, *133*, 42–51. [CrossRef]
50. Spielman, D.; Brook, B.W.; Frankham, R. Most species are not driven to extinction before genetic factors impact them. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15261–15264. [CrossRef]
51. Baker, A.J.; Pereira, S.L.; Paton, T.A. Phylogenetic relationships and divergence times of Charadriiformes genera: Multigene evidence for the Cretaceous origin of at least 14 clades of shorebirds. *Biol. Lett.* **2007**, *3*, 205–210. [CrossRef]
52. Wallis, G. *Genetic Status of New Zealand Black Stilt (*Himantopus novaezelandiae*) and Impact of Hybridisation*; Department of Conservation Wellington: Wellington, New Zealand, 1999.
53. Jarvis, E.D.; Mirarab, S.; Aberer, A.J.; Li, B.; Houde, P.; Li, C.; Ho, S.Y.; Faircloth, B.C.; Nabholz, B.; Howard, J.T.; et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **2014**, *346*, 1320–1331. [CrossRef]
54. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed on 7 November 2018).
55. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
56. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
57. Smeds, L.; Küstner, A. ConDeTri—a content dependent read trimmer for Illumina data. *PLoS ONE* **2011**, *6*, e26314. [CrossRef] [PubMed]
58. Simpson, J.T. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* **2014**, *30*, 1228–1235. [CrossRef]
59. Chikhi, R.; Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **2013**, *30*, 31–37. [CrossRef] [PubMed]
60. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* **2012**, *1*, 18. [CrossRef] [PubMed]
61. Bradnam, K.R.; Fass, J.N.; Alexandrov, A.; Baranay, P.; Bechner, M.; Birol, I.; Boisvert, S.; Chapman, J.A.; Chapuis, G.; Chikhi, R.; et al. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2013**, *2*, 10. [CrossRef]
62. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]

63. Waterhouse, R.M.; Seppey, M.; Simão, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **2017**, *35*, 543–548. [CrossRef]
64. Zdobnov, E.M.; Tegenfeldt, F.; Kuznetsov, D.; Waterhouse, R.M.; Simao, F.A.; Ioannidis, P.; Seppey, M.; Loetscher, A.; Kriventseva, E.V. OrthoDB v9. 1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **2016**, *45*, D744–D749. [CrossRef]
65. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644. [CrossRef] [PubMed]
66. Jiang, H.; Lei, R.; Ding, S.-W.; Zhu, S. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinf.* **2014**, *15*, 182. [CrossRef] [PubMed]
67. Crusoe, M.R.; Alameldin, H.F.; Awad, S.; Boucher, E.; Caldwell, A.; Cartwright, R.; Charbonneau, A.; Constantinides, B.; Edvenson, G.; Fay, S.; et al. The khmer software package: Enabling efficient nucleotide sequence analysis. *F1000Research* **2015**, *4*. [CrossRef] [PubMed]
68. Zerbino, D.; Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829. [CrossRef] [PubMed]
69. Goltsman, E.; Ho, I.; Rokhsar, D. Meraculous-2D: Haplotype-sensitive Assembly of Highly Heterozygous genomes. *arXiv* **2017**, arXiv:1703.09852.
70. Chapman, J.A.; Ho, I.Y.; Goltsman, E.; Rokhsar, D.S. Meraculous2: Fast accurate short-read assembly of large polymorphic genomes. *arXiv* **2016**, arXiv:1608.01031.
71. Kielbasa, S.M.; Wan, R.; Sato, K.; Horton, P.; Frith, M. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **2011**, *21*, 487–493. [CrossRef]
72. Tamazian, G.; Dobrynin, P.; Krashennikova, K.; Komissarov, A.; Koepfli, K.-P.; O'Brien, S.J. Chromosomer: A reference-based genome arrangement tool for producing draft chromosome sequences. *GigaScience* **2016**, *5*, 38. [CrossRef]
73. GigaDB. Genomic Data of the Killdeer (*Charadrius Vociferus*). 2014. Available online: <http://dx.doi.org/10.5524/101007> (accessed on 7 November 2018).
74. Gnerre, S.; MacCallum, I.; Przybylski, D.; Ribeiro, F.J.; Burton, J.N.; Walker, B.J.; Sharpe, T.; Hall, G.; Shea, T.P.; Sykes, S.; et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1513–1518. [CrossRef]
75. Ribeiro, F.; Przybylski, D.; Yin, S.; Sharpe, T.; Gnerre, S.; Abouelleil, A.; Berlin, A.M.; Montmayeur, A.; Shea, T.P.; Walker, B.J.; et al. Finished bacterial genomes from shotgun sequence data. *Genome Res.* **2012**, *22*, 2270–2277. [CrossRef]
76. Moraga, R. SemHelpers [Custom Perl Script]. 2017. Available online: <https://github.com/Lanilen/SemHelpers> (accessed on 7 November 2018).
77. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **2011**, *6*, e19379. [CrossRef] [PubMed]
78. Leigh, D.; Lischer, H.; Grossen, C.; Keller, L. Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. *Mol. Ecol. Resour.* **2018**. [CrossRef] [PubMed]
79. Murray, K.D.; Borevitz, J.O. Axe: Rapid, competitive sequence read demultiplexing using a trie. *Bioinformatics* **2017**. [CrossRef] [PubMed]
80. Herten, K.; Hestand, M.S.; Vermeesch, J.R.; Van Houdt, J.K. GBSX: A toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinform.* **2015**, *16*, 73. [CrossRef] [PubMed]
81. Moraga, R. Mux Barcodes [Custom Perl Script]. 2017. Available online: [https://github.com/sgalla32/mux\\_barcodes](https://github.com/sgalla32/mux_barcodes) (accessed on 7 November 2018).
82. Glaubitz, J.C.; Casstevens, T.M.; Lu, F.; Harriman, J.; Elshire, R.J.; Sun, Q.; Buckler, E.S. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **2014**, *9*, e90346. [CrossRef] [PubMed]
83. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357. [CrossRef]
84. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]

85. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
86. Moraga, R. Pancompare. 2018. Available online: <https://github.com/Lanilen/pancompare> (accessed on 7 November 2018).
87. Moraga, R. Split\_bamfile\_tasks.pl [Custom Perl Script]. 2018. Available online: <https://github.com/Lanilen/pancompare> (accessed on 7 November 2018).
88. Dodds, K.G.; McEwan, J.C.; Brauning, R.; Anderson, R.M.; Stijn, T.C.; Kristjánsson, T.; Clarke, S.M. Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genom.* **2015**, *16*, 1047. [CrossRef]
89. Bates, D.; Maechler, M.; Bolker, B.; Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
90. Gregory, T.R. Animal Genome Size Database. 2001. Available online: <http://www.genomesize.com/> (accessed on 7 November 2018).
91. Tigano, A.; Sackton, T.B.; Friesen, V.L. Assembly and RNA-free annotation of highly heterozygous genomes: The case of the thick-billed murre (*Uria lomvia*). *Mol. Ecol. Res.* **2018**, *18*, 79–90. [CrossRef] [PubMed]
92. Lacy, R.C.; Ballou, J.D.; Pollak, J.P. PMx: Software package for demographic and genetic analysis and management of pedigreed populations. *Methods Ecol. Evol.* **2012**, *3*, 433–437. [CrossRef]
93. Willoughby, J.R.; Fernandez, N.B.; Lamb, M.C.; Ivy, J.A.; Lacy, R.C.; DeWoody, J.A. The impacts of inbreeding, drift and selection on genetic diversity in captive breeding populations. *Mol. Ecol.* **2015**, *24*, 98–110. [CrossRef]
94. Putnam, A.S.; Ivy, J.A. Kinship-based management strategies for captive breeding programs when pedigrees are unknown or uncertain. *J. Hered.* **2013**, *105*, 303–311. [CrossRef] [PubMed]
95. Hammerly, S.C.; de la Cerda, D.A.; Bailey, H.; Johnson, J.A. A pedigree gone bad: Increased offspring survival after using DNA-based relatedness to minimize inbreeding in a captive population. *Anim. Conserv.* **2016**, *19*, 296–303. [CrossRef]
96. Szulkin, M.; Bierne, N.; David, P. Heterozygosity-fitness correlations: A time for reappraisal. *Evolution* **2010**, *64*, 1202–1217. [CrossRef] [PubMed]
97. Sandoval-Castillo, J.; Attard, C.R.M.; Marri, S.; Brauer, C.J.; Möller, L.M.; Beheregaray, L.B. A user-friendly computer program to establish captive breeding groups that minimize relatedness without pedigree information. *Mol. Ecol. Res.* **2017**, *17*, 278–287. [CrossRef]
98. Trapnell, C.; Salzberg, S.L. How to map billions of short reads onto genomes. *Nat. Biotechnol.* **2009**, *27*, 455. [CrossRef]
99. Kajitani, R.; Toshimoto, K.; Noguchi, H.; Toyoda, A.; Ogura, Y.; Okuno, M.; Yabana, M.; Harada, M.; Nagayasu, E.; Maruyama, H.; et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **2014**. [CrossRef]
100. IUCN. The IUCN Red List of Threatened Species. 2018. Available online: <http://www.iucnredlist.org> (accessed on 7 November 2018).
101. Küpper, C.; Stocks, M.; Risse, J.E.; dos Remedios, N.; Farrell, L.L.; McRae, S.B.; Morgan, T.C.; Karlionova, N.; Pinchuk, P.; Verkuil, Y.I.; et al. A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **2016**, *48*, 79. [CrossRef]
102. Frankham, R. Challenges and opportunities of genetic approaches to biological conservation. *Biol. Conserv.* **2010**, *143*, 1919–1927. [CrossRef]
103. Taylor, H.R. The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecol. Evol.* **2015**, *5*, 3140–3150. [CrossRef] [PubMed]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

# Appendix E: Chapter 4, Supplemental Materials

## E1: Samples

Table E 1: Samples used in genotyping-by-sequencing and resequencing analyses.

Sample ID	Tissue Type	Year Collected	GBS - Batch 1	GBS - Batch 2	Resequencing	Description
240	Blood	2015	✓		✓	Captive Parent
451	Blood	2015	✓		✓	Captive Parent
452	Blood	2015	✓	✓	✓	Captive Parent
453	Blood	2015	✓	✓	✓	Captive Parent
639	Blood	2015	✓	✓	✓	Captive Parent
1360	Blood	2015	✓			Captive Parent
1376	Blood	2015	✓	✓	✓	Captive Parent
1377	Blood	2015	✓	✓	✓	Captive Parent
1409	Blood	2015	✓			Captive Parent
1429	Blood	2015	✓	✓	✓	Captive Parent
1469	Blood	2015	✓	✓	✓	Captive Parent
1565	Blood	2015	✓	✓	✓	Captive Parent
1659	Blood	2015	✓	✓	✓	Captive Parent
1661	Blood	2015	✓	✓	✓	Captive Parent
1736	Blood	2015	✓			Wild Offspring
1738	Blood	2015	✓		✓	Wild Offspring
1744	Blood	2015	✓			Wild Offspring
1762	Blood	2015	✓			Wild Offspring
1763	Blood	2015	✓			Wild Offspring
1764	Blood	2015	✓			Wild Offspring
1860	Blood	2015	✓			Wild Offspring
1864	Blood	2015	✓			Wild Offspring
1872	Blood	2015	✓	✓	✓	Wild Offspring
1892	Blood	2016			✓	Wild Offspring
1903	Blood	2016	✓		✓	Captive Offspring
1904	Blood	2016	✓			Captive Offspring
1905	Blood	2016	✓			Captive Offspring
1906	Blood	2016	✓			Captive Offspring
1907	Blood	2016	✓			Captive Offspring
1908	Blood	2016	✓			Captive Offspring
1915	Blood	2016	✓			Captive Offspring
1920	Blood	2016	✓			Captive Offspring
1921	Blood	2016	✓		✓	Captive Offspring
1931	Blood	2016	✓			Captive Offspring
1932	Blood	2016	✓			Captive Offspring
1934	Blood	2016			✓	Wild Offspring
1936	Blood	2016			✓	Wild Offspring
1944	Blood	2016	✓			Captive Offspring
1947	Blood	2016	✓			Captive Offspring
1948	Blood	2016	✓			Captive Offspring
1949	Blood	2016	✓			Captive Offspring

1950	Blood	2016	✓			Captive Offspring
1952	Blood	2016	✓			Captive Offspring
1953	Blood	2016	✓			Captive Offspring
1957	Blood	2016	✓			Captive Offspring
1964	Blood	2016	✓			Captive Offspring
1966	Blood	2016	✓		✓	Captive Offspring
1969	Blood	2016	✓			Captive Offspring
1970	Blood	2016	✓			Captive Offspring
1971	Blood	2016	✓			Captive Offspring
1975	Blood	2016	✓		✓	Captive Offspring
1976	Blood	2016	✓			Captive Offspring
15_128	Muscle	2016	✓			Captive Offspring
15_129	Muscle	2016	✓			Captive Offspring
15_130	Muscle	2016	✓			Captive Offspring
1980	Blood	2017			✓	Wild Offspring
1993	Blood	2017		✓		Captive Offspring
1994	Blood	2017		✓		Captive Offspring
1995	Blood	2017		✓		Captive Offspring
1996	Blood	2017		✓		Captive Offspring
1998	Blood	2017		✓		Captive Offspring
2000	Blood	2017		✓	✓	Captive Offspring
2004	Blood	2017		✓	✓	Captive Offspring
2012	Blood	2017			✓	Wild Offspring
2013	Blood	2017		✓		Captive Offspring
2014	Blood	2017		✓		Captive Offspring
2015	Blood	2017		✓		Captive Offspring
2016	Blood	2017		✓	✓	Captive Offspring
2023	Blood	2017			✓	Wild Offspring
2028	Blood	2017		✓		Captive Offspring
2029	Blood	2017		✓		Captive Offspring
2030	Blood	2017		✓		Captive Offspring
2032	Blood	2017			✓	Wild Offspring
2034	Blood	2017		✓		Captive Offspring
2035	Blood	2017			✓	Wild Offspring
2050	Blood	2017		✓		Captive Offspring
2054	Blood	2017		✓		Captive Offspring
2057	Blood	2017		✓		Captive Offspring
2058	Blood	2017		✓		Captive Offspring
2059	Blood	2017		✓		Captive Offspring
2074	Blood	2017			✓	Wild Offspring
2078	Blood	2017		✓		Captive Offspring
2081	Blood	2017		✓	✓	Captive Offspring
2091	Blood	2017			✓	Wild Offspring
2110	Blood	2017		✓		Captive Offspring
2111	Blood	2017		✓		Captive Offspring
2112	Blood	2017		✓		Captive Offspring
2115	Blood	2017		✓		Captive Offspring
2118	Blood	2017		✓		Captive Offspring
2119	Blood	2017		✓		Captive Offspring
2120	Blood	2017		✓		Captive Offspring
2121	Blood	2017		✓	✓	Captive Offspring

2122	Blood	2017		✓		Captive Offspring
2123	Blood	2017		✓		Captive Offspring
2124	Blood	2017		✓	✓	Captive Offspring
2125	Blood	2017		✓	✓	Captive Offspring
15_114	Germinal Disk	2017		✓	✓	Captive Offspring
16_05	Germinal Disk	2017		✓		Captive Offspring

## E2: Sequencing Costs

*Table E 2: Cost associated with genome sequencing and alignment. Prices for sequencing from competitive sequencing providers (see <https://genohub.com/ngs/>). Compute instance and storage costs from competitive New Zealand-based cloud computing services (see <https://catalystcloud.nz>). Work hours computed from bioinformatician analysis time at €58 per hour*

Genome Type	Genome Description	Sequencing	Compute Instance and Storage	Work hours	TOTAL COST
Basic Genome	1 lane of Illumina sequencing; 5 days of 256 Gb memory instance for initial assembly plus refinement; 2Tb temporary storage; 60 hours of bioinformatician analysis time.	€ 1,750	€ 315	€ 3,500	€ 5,565
Short Read Full Genome	100x coverage pair-end reads and 2 mate-pair libraries; 5 days of 256 Gb memory instance for pair-end assembly and 9 days of 128 Gb instance for multiple scaffolding iterations; 2 Tb temporary storage; 100 hours of bioinformatician analysis time.	€ 9,335	€ 370	€ 5,835	€ 15,540
Multi-tech Genome	100x coverage pair-end reads, multiple mate-pair libraries, and 10x coverage PacBio sequencing; 2 Tb temp storage. 2 days with 6 instances of 32 Gb machines for PacBio error correction and 2 days with 6 instances for gap filling/completion with PacBio; 160h of bioinformatician analysis time.	€ 17,500	€ 875	€ 9,335	€ 27,710
Long-read Genome	50x PacBio sequencing, 1 lane of Illumina sequencing, and 5 flowcells of Oxford Nanopore; 14 days of 256 Gb instance for long read assemblies, followed by 5 days of 256Gb instance for short read assemblies, and 5 days of 128 Gb instance for integration of all three technologies; 3Tb of storage; 180 hours of bioinformatician analysis time.	€ 46,670	€ 1,460	€ 10,500	€ 58,630

## E3: Lithium Chloride DNA Extractions Protocol

A small piece of dried blood or tissue was isolated for each sample. 250µl of lysis buffer (1.0M tris, 0.5M EDTA, 5.0M NaCl, 10% SDS, 50mg/mL proteinase K, and molecular-grade RNase-free water) was added and the reaction was placed on a 55°C heat block overnight (at least 6 hours). When fully digested, 250µl of 5M LiCl solution was added to the digested blood or tissue. Samples were inverted for 1 minute. In a fume hood, 500µl of 24:1 chloroform:isoamyl alcohol was added to each tube. Samples were inverted for 5 minutes. Samples were then centrifuged at 11,000 RPM for 3 minutes, where samples separate into two phases.

The complete top phase of sample was transferred into a new tube and 250µl of 5M LiCl solution was added. Samples were inverted for 1 minute. In a fume hood, 500µl of 24:1 chloroform:isoamyl alcohol was added to each tube. Samples were mixed by inverting for 2 minutes. Samples were then centrifuged at 11,000 rpm for 1 minute, where samples separate into two phases. 500µl of the top phase was transferred to a new 1.5ml tube, being careful not to touch the intermediate film between the two phases. 1ml of cold (-20°C) absolute ethanol was added to each sample. Samples were inverted gently as DNA precipitates out of solution. Samples were then placed in the -20°C

freezer for a minimum of 2 hours. Once out of the freezer, tubes were centrifuged for 10 minutes at 12,000 rpm at -20°C.

The supernate in each sample was poured off into a waste container, being careful not to disturb the pellet. 500µl of 70% ethanol was added to each sample to wash the pellet. The pellet was fished out using a pipette tip and transferred to a clean tube. After the pellet air dried, the DNA was resuspended in 50µL of TE8 buffer. The samples were incubated for 37°C for 40 minutes to allow for improved re-suspension. Samples were then stored at 4°C until used, the permanently at -20°C.

# Appendix F: Chapter 4, SNP Discovery Pipelines and R-code

## F.1: GBS SNP discovery pipelines

These scripts are used for processing the 2016 and 2018 GBS datasets.

1. `Axe_demux_2016_2018.sh` was used to trim and demultiplex reads.
2. `Fastqc_GBS_Data_Raw2016_2018.sh` was used to run FastQC on all raw reads.
3. `Skewer_GBS_2016_2018.sh` was used to trim reads.
4. `Velvet_GBS_2016_2018.sh` was used to interleave forward and reverse reads.
5. `GBSX_MuxBarcodes.sh` was used to generate barcodes and append them to interleaved reads.
6. Tassel5 shell scripts was used to discover SNPs.
7. `VCFTools_BCFTools_GBS.sh` was used to filter SNPs.

### F.1.1: Axe

Available at: [https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/axe-demux\\_2016\\_2018.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/axe-demux_2016_2018.sh)

```
#!/bin/sh

axe-demux -b '/home/stephanie/Desktop/GBS Raw
Reads/Raw_Reads_2016/ApeK1PlatelKakiKey_Repaired.txt' -c -f
'/home/stephanie/Desktop/GBS Raw Reads/Raw_Reads_2016/2264-L1-24-
1_S1_L001_R1_001.fastq.gz' -r '/home/stephanie/Desktop/GBS Raw
Reads/Raw_Reads_2016/2264-L1-24-1_S1_L001_R2_001.fastq.gz' -F
'/home/stephanie/Desktop/GBS Raw Reads/Raw_Reads_2016/Forward' -R
'/home/stephanie/Desktop/GBS Raw Reads/Raw_Reads_2016/Demultiplexed_Reads/Reverse'
/

axe-demux -b
'/media/stephanie/External4/Raw_Reads_2018/30001_Key_IDs_kaki_only.txt' -c -f
'/media/stephanie/External4/Raw_Reads_2018/CS121_R1.fq.gz' -r
'/media/stephanie/External4/Raw_Reads_2018/CS121_R2.fq.gz' -F
'/media/stephanie/External4/Raw_Reads_2018/Forward' -R
'/media/stephanie/External4/Raw_Reads_2018/Reverse'
```

### F.1.2: FastQC v. 0.11.4

Found at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/Fastqc\\_GBS\\_Data\\_Raw2016\\_2018.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/Fastqc_GBS_Data_Raw2016_2018.sh)

```
#!/bin/sh

perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2016_GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/
Demultiplexed/15_128D_R1.fastq'\
```



[illegible]



[illegible]





















```

perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/142_
9_R1.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137_
7_R2.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137_
7_R1.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137_
6_R2.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137_
6_R1.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/639_
R2.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/639_
R1.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/453_
R2.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/453_
R1.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/452_
R2.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/452_
R1.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/16_
05_R2.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/16_
05_R1.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/15_
114_R2.fastq' \
perl '/home/stephanie/FastQC/fastqc'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/15_
114_R1.fastq' \

```

### F.1.3: Skewer

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/Skewer\\_GBS\\_2016\\_2018.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/Skewer_GBS_2016_2018.sh)

```
#!/bin/sh
```

```

skewer -l 32 -Q 30 -n
'/media/stephanie/Data/GBS/2016_GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/
Demultiplexed/15_128D_R1.fastq'
'/media/stephanie/Data/GBS/2016_GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/
Demultiplexed/15_128D_R2.fastq' -o
'/media/stephanie/Data/GBS/2016_GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/
Demultiplexed/15_128D_skewer_trimmed' /

```

```

skewer -l 32 -Q 30 -n
'/media/stephanie/Data/GBS/2016_GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/
Demultiplexed/15_128_R1.fastq'

```



























```

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/16_
05_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/16_
05_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/16_
05_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/452
_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/452
_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/452
_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/453
_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/453
_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/453
_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/639
_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/639
_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/639
_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137
6_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137
6_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137
6_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137
7_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137
7_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/137
7_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/142
9_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/142
9_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/142
9_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/146
9_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/146
9_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/146
9_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/156
5_R1.fastq'

```









```

'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
0_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
0_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
1_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
1_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
1_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
2_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
2_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
2_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
3_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
3_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
3_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
4_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
4_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
4_skewer_trimmed_100bp' /

skewer -l 32 -L 100 -e -Q 30 -n
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
5_R1.fastq'
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
5_R2.fastq' -o
'/media/stephanie/Data/GBS/2018_GBS/Demultiplexed_Fastq/Demultiplexed_Fastqs/212
5_skewer_trimmed_100bp' /

```

#### F.1.4: Velvet

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/Velvet\\_GBS\\_2016\\_2018.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/Velvet_GBS_2016_2018.sh)

```
#!/bin/sh
```

```

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/15_114_skewer_trimme
d_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/15_114_skewer_trimme
d_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/15_114_100BP_Quality
30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/16_05_skewer_trimmed
_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/16_05_skewer_trimmed

```

```

_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/16_05_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/452_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/452_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/452_2018_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/453_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/453_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/453_2018_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/639_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/639_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/639_2018_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1376_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1376_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1376_2018_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1377_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1377_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1377_2018_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1429_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1429_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1429_2018_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1469_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1469_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1469_2018_100BP_Quality30_n_interleaved.fastq /

perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1565_skewer_trimmed_100bp-pair1.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1565_skewer_trimmed_100bp-pair2.fastq
/media/stephanie/Data/GBS/2018_GBS/Trimmed100BP_Quality30_n/1565_2018_100BP_Quality30_n_interleaved.fastq /

```

























```
perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmmed_Quality30_n/1872D_skewer_trimmed-trimmed-pair1.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmmed_Quality30_n/1872D_skewer_trimmed-trimmed-pair2.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmmed_Quality30_n/1872D_interleaved_Quality30_n.fastq /
```









```
mmed_Quality30_n/1971D_skewer_trimmed-trimmed-pair2.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1971D_interleaved_Quality30_n.fastq /
```

```
perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1971_skewer_trimmed-trimmed-pair1.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1971_skewer_trimmed-trimmed-pair2.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1971_interleaved_Quality30_n.fastq /
```

```
perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1975_skewer_trimmed-trimmed-pair1.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1975_skewer_trimmed-trimmed-pair2.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1975_interleaved_Quality30_n.fastq /
```

```
perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1976D_skewer_trimmed-trimmed-pair1.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1976D_skewer_trimmed-trimmed-pair2.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1976D_interleaved_Quality30_n.fastq /
```

```
perl /home/stephanie/Desktop/shuffleSequences_fastq.pl
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1976_skewer_trimmed-trimmed-pair1.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1976_skewer_trimmed-trimmed-pair2.fastq
/media/stephanie/Data/GBS/Demultiplexed_Fastqs_Paired_With_100bp_2018/Skewer_Tri
mmed_Quality30_n/1976_interleaved_Quality30_n.fastq /
```

### ***F.1.5: GBSX***

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/GBSX\\_MuxBarcodes.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/GBSX_MuxBarcodes.sh)

```
#Used this for unique barcode genertor
Java -jar "location_of_GBSX_Releasess/GBSX_v1.0.jar" --BarcodeGenerator -b 141 -
e ApeKI > 141barcodes.txt
```

```
#Place 141barcodes.txt in a folder with all interleaved fastq files.
Perl mux_barcodes.pl 96barcodes.txt ./
```

### ***F.1.6: Tassel***

#### **F.1.6.1: Kaki**

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/Tassel5\\_Kaki.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/Tassel5_Kaki.sh)

```
#!/bin/sh
```

```
'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-
TEGenzymes.v2/tassel5-src/run_pipeline.pl' -Xmx16G -fork1 -GBSSeqToTagDBPlugin -
e ApeKI -i '/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/2016_2018_Trimmed_ConcatSamples_Kaki.db' -k
```

```

'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode
d_concat_samples.txt' -kmerLength 64 -minKmerL 20 -mnQS 30 -mxKmerNum 100000000 -
endPlugin -runfork1

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-
TEGenzymes.v2/tassel5-src/run_pipeline.pl' -Xmx16G -fork1 -
TagExportToFastqPlugin -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/2016_2018_Trimmed_ConcatSamples_Kaki.db' -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/tagsForAlign.fa.gz' -c 10 -endPlugin -runfork1

bowtie2 -p 15 --very-sensitive -x
'/media/stephanie/Data/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed
/Kaki1_v2.2_concat_index.fasta' -U
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/tagsForAlign.fa.gz' -S
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/tagsForAlign.sam'

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-
TEGenzymes.v2/tassel5-src/run_pipeline.pl' -Xmx16G -fork1 -SAMToGBSdbPlugin -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/tagsForAlign.sam' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/2016_2018_Trimmed_ConcatSamples_Kaki.db' -aProp 0.0 -aLen 0 -endPlugin -
runfork1

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-
TEGenzymes.v2/tassel5-src/run_pipeline.pl' -Xmx16G -fork1 -
DiscoverySNPCallerPluginV2 -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/2016_2018_Trimmed_ConcatSamples_Kaki.db' -sC 1 -eC 1 -mnLCov 0.1 -mnMAF
0.05 -deleteOldData true -endPlugin -runfork1

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-
TEGenzymes.v2/tassel5-src/run_pipeline.pl' -Xmx16G -fork1 -
ProductionSNPCallerPluginV2 -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/2016_2018_Trimmed_ConcatSamples_Kaki.db' -e ApeKI -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -k
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode
d_concat_samples.txt' -kmerLength 64 -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSa
mples/Kaki_5Sept2018.vcf' -endPlugin -runfork1

```

### F.1.6.2: Pied Stilt

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/Tassel5\\_PiedStilt.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/Tassel5_PiedStilt.sh)

```
#!/bin/sh
```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -GBSSeqToTagDBPlugin -e ApeKI -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/2016_2018_Trimmed_ConcatSamples_AusPied.db' -k
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode
d_concat_sa
mples.txt' -kmerLength 64 -minKmerL 20 -mnQS 30 -mxKmerNum 100000000 -endPlugin -
runfork1

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -TagExportToFastqPlugin -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat

```



```

Samples/2016_2018_Trimmed_ConcatSamples_AusPied.db' -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/tagsForAlign.fa.gz' -c 10 -endPlugin -runfork1

bowtie2 -p 15 --very-sensitive -x
'/media/stephanie/Data/Genomes/AusPiedGenomes/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_ind
exed.fasta' -U
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/tagsForAlign.fa.gz' -S
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/tagsForAlign.sam'

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -SAMToGBSdbPlugin -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/tagsForAlign.sam' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/2016_2018_Trimmed_ConcatSamples_AusPied.db' -aProp 0.0 -aLen 0 -endPlugin -
runfork1

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -DiscoverySNPCallerPluginV2 -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/2016_2018_Trimmed_ConcatSamples_AusPied.db' -sC 1 -eC 1 -mnLCov 0.1 -mnMAF
0.05 -deleteOldData true -endPlugin -runfork1

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -ProductionSNPCallerPluginV2 -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/2016_2018_Trimmed_ConcatSamples_AusPied.db' -e ApeKI -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -k
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode_concat_sa
mples.txt' -kmerLength 64 -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/PiedStilt_17Aug2018_Concat
Samples/AusPied_17Aug2018.vcf' -endPlugin -runfork1

```

### F.1.6.3: Pied Avocet

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/Tassel5\\_Avocet.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/Tassel5_Avocet.sh)

```
#!/bin/sh
```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -GBSSeqToTagDBPlugin -e ApeKI -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/2016_2018_Trimmed_ConcatSamples_Kaki.db' -k
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode_concat_sa
mples.txt' -kmerLength 64 -minKmerL 20 -mnQS 30 -mxKmerNum 100000000 -endPlugin -
runfork1

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -TagExportToFastqPlugin -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/2016_2018_Trimmed_ConcatSamples_Kaki.db' -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/tagsForAlign.fa.gz' -c 10 -endPlugin -runfork1

bowtie2 -p 15 --very-sensitive -x
'/media/stephanie/Data/Genomes/Kaki_Superscaffolds/Kaki_v2_Concat/Bowtie_Indexed/Ka
kil_v2.2_concat_index.fasta' -U
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/tagsForAlign.fa.gz' -S
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/tagsForAlign.sam'

```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -SAMToGBSdbPlugin -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/tagsForAlign.sam' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/2016_2018_Trimmed_ConcatSamples_Kaki.db' -aProp 0.0 -aLen 0 -endPlugin -
runfork1

```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -DiscoverySNPCallerPluginV2 -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/2016_2018_Trimmed_ConcatSamples_Kaki.db' -sC 1 -eC 1 -mnLCov 0.1 -mnMAF 0.05 -
deleteOldData true -endPlugin -runfork1

```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -ProductionSNPCallerPluginV2 -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/2016_2018_Trimmed_ConcatSamples_Kaki.db' -e ApeKI -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -k
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode_concat_sa
mples.txt' -kmerLength 64 -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Kaki_5Sept2018_ConcatSampl
es/Kaki_5Sept2018.vcf' -endPlugin -runfork1

```

#### F.1.6.4: Killdeer

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/Tassel5\\_Killdeer.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/Tassel5_Killdeer.sh)

```
#!/bin/sh
```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -GBSSeqToTagDBPlugin -e ApeKI -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatS
amples/2016_2018_Trimmed_ConcatSamples_Killdeer.db' -k
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode_concat_sa
mples.txt' -kmerLength 64 -minKmerL 20 -mnQS 30 -mxKmerNum 100000000 -endPlugin -
runfork1

```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -TagExportToFastqPlugin -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatS
amples/2016_2018_Trimmed_ConcatSamples_Killdeer.db' -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatS
amples/tagsForAlign.fa.gz' -c 10 -endPlugin -runfork1

```

```

bowtie2 -p 15 --very-sensitive -x
'/media/stephanie/Data/Genomes/Killdeer/Killdeer_Concat/Bowtie/killdeer_concat_inde
xed.fasta' -U
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatS
amples/tagsForAlign.fa.gz' -S
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatS
amples/tagsForAlign.sam'

```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -SAMToGBSdbPlugin -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatS
amples/tagsForAlign.sam' -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatS
amples/2016_2018_Trimmed_ConcatSamples_Killdeer.db' -aProp 0.0 -aLen 0 -endPlugin -
runfork1

```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-
src/run_pipeline.pl' -Xmx16G -fork1 -DiscoverySNPCallerPluginV2 -db

```

```

'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatSamples/2016_2018_Trimmed_ConcatSamples_Killdeer.db' -sC 1 -eC 1 -mnLCov 0.1 -mnMAF 0.05 -deleteOldData true -endPlugin -runfork1

```

```

'/home/stephanie/Desktop/Rogers_Tassel_Modifications/tassel5-TEGenzymes.v2/tassel5-src/run_pipeline.pl' -Xmx16G -fork1 -ProductionSNPCallerPluginV2 -db
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatSamples/2016_2018_Trimmed_ConcatSamples_Killdeer.db' -e ApeKI -i
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/' -k
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/keyfile_barcode_concat_samples.txt' -kmerLength 64 -o
'/media/stephanie/Data/GBS/Tassel5/Tassel5_14Aug2018_30Q/Killdeer_21Aug2018_ConcatSamples/AusPied_17Aug2018.vcf' -endPlugin -runfork1 1

```

### ***F.1.7: VCFTools & BCFTools v. 1.9***

Available at:

[https://github.com/sgalla32/GBS\\_2016\\_2018\\_Data\\_Prep/blob/master/VCFTools\\_BCFTools\\_GBS.sh](https://github.com/sgalla32/GBS_2016_2018_Data_Prep/blob/master/VCFTools_BCFTools_GBS.sh)

```
#!/bin/sh
```

```
#Kaki GBS vcf filtering.
```

```
vcftools --vcf Kaki_5Sept2018_RmvLowDepth.recode.vcf --min-alleles 2 --max-alleles 2 --min-meanDP 5 --max-missing 0.9 --out
Kaki_5Sept2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9 --recode
```

```
bcftools +prune -l 0.8 -w 1000
Kaki_5Sept2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9.recode.vcf -Ov -o
Kaki_5Sept2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf
```

```
vcftools --vcf
Kaki_5Sept2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf --thin 65 --out Kaki_5Sept2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000_thin65 --recode
```

```
#AusPied GBS vcf filtering.
```

```
vcftools --vcf AusPied_17Aug2018_RmvLowDepth.recode.vcf --min-alleles 2 --max-alleles 2 --min-meanDP 5 --max-missing 0.9 --out
AusPied_17Aug2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9 --recode
```

```
bcftools +prune -l 0.8 -w 1000
AusPied_17Aug2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9.recode.vcf -Ov -o
AusPied_17Aug2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf
```

```
vcftools --vcf
AusPied_17Aug2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf --thin 65 --out
AusPied_17Aug2018_RmvLowDp_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000_thin65 --recode
```

```
#Avocet GBS vcf filtering.
```

```
vcftools --vcf Avocet_20Aug2018_RmvLowDepth.recode.vcf --min-alleles 2 --max-alleles 2 --min-meanDP 5 --max-missing 0.9 --out
Avocet_20Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9 --recode
```

```
bcftools +prune -l 0.8 -w 1000
Avocet_20Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9.recode.vcf -Ov -o
Avocet_20Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf
```

```
vcftools --vcf
Avocet_20Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf --thin
```

```

65 --out
Avocet_20Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000_thin65 --
recode

#Killdeer GBS vcf filtering.

vcftools --vcf Killdeer_21Aug2018_RmvLowDepth.recode.vcf --min-alleles 2 --max-
alleles 2 --min-meanDP 5 --max-missing 0.9 --out
Killdeer_21Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9 --recode

bcftools +prune -l 0.8 -w 1000
Killdeer_21Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9.recode.vcf -Ov -o
Killdeer_21Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf

vcftools --vcf
Killdeer_21Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000.vcf --
thin 65 --out
Killdeer_21Aug2018_RmvLowDepth_Biallelic_MinMeanDP5_Missing0.9_LD0.8_w1000_thin65 -
-recode

```

## F.2: Resequencing SNP-Discovery Pipelines

### F.2.1: FastQC v. 0.11.4

Available at:

[https://github.com/sgalla32/Resequence\\_Data\\_Processing/blob/master/FastQC\\_Resequencing.sh](https://github.com/sgalla32/Resequence_Data_Processing/blob/master/FastQC_Resequencing.sh)

```
#!/bin/sh
```

```

/share/data3/conda/miniconda/bin/fastqc /share/data3/sga94/IKMB_Data/fastq/H01383-
L1_S41_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01383-L1_S41_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01384-L1_S42_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01384-L1_S42_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01385-L1_S43_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01385-L1_S43_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01386-L1_S1_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01386-L1_S1_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01387-L1_S2_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01387-L1_S2_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01388-L1_S3_L002_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01388-L1_S3_L002_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01389-L1_S4_L003_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01389-L1_S4_L003_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01390-L1_S5_L003_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01390-L1_S5_L003_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01391-L1_S6_L004_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01391-L1_S6_L004_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01392-L1_S7_L005_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01392-L1_S7_L005_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01393-L1_S8_L005_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01393-L1_S8_L005_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01394-L1_S9_L005_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01394-L1_S9_L005_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01395-L1_S10_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01395-L1_S10_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01396-L1_S11_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01396-L1_S11_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01397-L1_S12_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01397-L1_S12_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01398-L1_S1_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01398-L1_S1_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01399-L1_S2_L001_R1_001.fq \

```

```

/share/data3/sga94/IKMB_Data/fastq/H01399-L1_S2_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01400-L1_S3_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01400-L1_S3_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01401-L1_S1_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01401-L1_S1_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01402-L1_S2_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01402-L1_S2_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01403-L1_S3_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01403-L1_S3_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01404-L1_S7_L005_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01405-L1_S8_L005_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01405-L1_S8_L005_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01406-L1_S9_L005_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01406-L1_S9_L005_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01407-L1_S35_L006_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01407-L1_S35_L006_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01408-L1_S36_L006_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01408-L1_S36_L006_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01409-L1_S37_L006_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01409-L1_S37_L006_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01410-L1_S38_L006_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01410-L1_S38_L006_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01411-L1_S39_L006_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01411-L1_S39_L006_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01412-L1_S40_L006_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01412-L1_S40_L006_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01413-L1_S10_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01413-L1_S10_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01414-L1_S11_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01414-L1_S11_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01415-L1_S12_L007_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01415-L1_S12_L007_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01416-L1_S1_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01416-L1_S1_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01417-L1_S2_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01417-L1_S2_L001_R2_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01418-L1_S3_L001_R1_001.fq \
/share/data3/sga94/IKMB_Data/fastq/H01418-L1_S3_L001_R2_001.fq

```

## F.2.2: Trimmomatic

Available at:

[https://github.com/sgalla32/Resequence\\_Data\\_Processing/blob/master/Trimmomatic.sh](https://github.com/sgalla32/Resequence_Data_Processing/blob/master/Trimmomatic.sh)

```
#!/bin/sh
```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01395-L1_S10_L007_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01395-L1_S10_L007_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R1_001_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01395-L1_S10_L008_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01395-L1_S10_L008_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L008_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-

```

```

L1_S10_L008_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L008_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L008_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01396-L1_S11_L007_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01396-L1_S11_L007_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01396-L1_S11_L008_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01396-L1_S11_L008_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L008_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L008_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L008_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L008_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01397-L1_S12_L007_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01397-L1_S12_L007_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01397-L1_S12_L008_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01397-L1_S12_L008_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L008_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L008_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L008_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L008_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01398-L1_S1_L001_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01398-L1_S1_L001_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R2_001_paired_trimmed.fq.gz

```

```
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-  
L1_S1_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10  
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33  
/share/data3/sga94/IKMB_Data/fastq/H01398-L1_S1_L002_R1_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/H01398-L1_S1_L002_R2_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-  
L1_S41_L008_R1_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-  
L1_S1_L002_R1_001_unpaired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-  
L1_S1_L002_R2_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-  
L1_S1_L002_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10  
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33  
/share/data3/sga94/IKMB_Data/fastq/H01399-L1_S2_L001_R1_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/H01399-L1_S2_L001_R2_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L001_R1_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L001_R1_001_unpaired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L001_R2_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10  
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33  
/share/data3/sga94/IKMB_Data/fastq/H01399-L1_S2_L002_R1_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/H01399-L1_S2_L002_R2_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L002_R1_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L002_R1_001_unpaired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L002_R2_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-  
L1_S2_L002_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10  
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33  
/share/data3/sga94/IKMB_Data/fastq/H01400-L1_S3_L001_R1_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/H01400-L1_S3_L001_R2_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L001_R1_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L001_R1_001_unpaired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L001_R2_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10  
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33  
/share/data3/sga94/IKMB_Data/fastq/H01400-L1_S3_L002_R1_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/H01400-L1_S3_L002_R2_001.fastq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L002_R1_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L002_R1_001_unpaired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L002_R2_001_paired_trimmed.fq.gz  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-  
L1_S3_L002_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10  
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01401-L1_S1_L001_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01401-L1_S1_L001_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01401-L1_S4_L003_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01401-L1_S4_L003_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S4_L003_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S4_L003_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S4_L003_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S4_L003_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01402-L1_S2_L001_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01402-L1_S2_L001_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01402-L1_S5_L003_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01402-L1_S5_L003_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S5_L003_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S5_L003_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S5_L003_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S5_L003_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01403-L1_S3_L001_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01403-L1_S3_L001_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01403-L1_S6_L003_R1_001.fastq.gz

```



```

/share/data3/sga94/IKMB_Data/fastq/H01403-L1_S6_L003_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S6_L003_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S6_L003_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S6_L003_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S6_L003_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01404-L1_S7_L005_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01404-L1_S7_L005_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01404-L1_S7_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01404-L1_S7_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01405-L1_S8_L005_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01405-L1_S8_L005_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01405-L1_S8_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01405-L1_S8_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

```

```

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01406-L1_S9_L005_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01406-L1_S9_L005_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R1_001_paired_trimmed.fq.gz

```

```

/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomaticH01406-
L1_S9_L005_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01406-L1_S9_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01406-L1_S9_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01407-L1_S35_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01407-L1_S35_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01408-L1_S36_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01408-L1_S36_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01409-L1_S37_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01409-L1_S37_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \

/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01410-L1_S38_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01410-L1_S38_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-

```

```
L1_S38_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01411-L1_S39_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01411-L1_S39_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01412-L1_S40_L006_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01412-L1_S40_L006_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01413-L1_S10_L007_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01413-L1_S10_L007_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01414-L1_S11_L007_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01414-L1_S11_L007_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01415-L1_S12_L007_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01415-L1_S12_L007_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
```

```
L1_S12_L007_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01416-L1_S1_L001_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01416-L1_S1_L001_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

```
/share/data3/conda/miniconda3/bin/trimmomatic PE -threads 32 -phred33
/share/data3/sga94/IKMB_Data/fastq/H01417-L1_S2_L001_R1_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/H01417-L1_S2_L001_R2_001.fastq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R1_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R1_001_unpaired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R2_001_paired_trimmed.fq.gz
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R2_001_unpaired_trimmed.fastq.gz ILLUMINACLIP:TruSeq2-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50 \
```

## F.2.3: Bowtie2

### F.2.3.1: Kaki

Available at:

[https://github.com/sgalla32/Resequence\\_Data\\_Processing/blob/master/Bowtie2\\_Kaki.sh](https://github.com/sgalla32/Resequence_Data_Processing/blob/master/Bowtie2_Kaki.sh)

```
#!/bin/sh
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
L1_S43_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
```

```

L1_S43_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-
L1_S3_L002_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-
L1_S3_L002_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-
L1_S4_L003_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-
L1_S4_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-
L1_S5_L003_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-
L1_S5_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-
L1_S6_L004_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-
L1_S6_L004_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-

```

```

L1_S7_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-
L1_S7_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-
L1_S8_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-
L1_S8_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-
L1_S9_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-
L1_S9_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_

```

```

v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kakil_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405_Concat_1set_kaki.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406_Concat_1set_kaki.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-L1_S35_L006.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-L1_S36_L006.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-L1_S37_L006.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-L1_S38_L006.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-L1_S39_L006.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |

```



```

/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-L1_S40_L006.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-L1_S10_L007.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-L1_S11_L007.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-L1_S12_L007.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-L1_S1_L001.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-L1_S2_L001.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Kaki_SuperScaffolds/Kaki_v2_Concat/Bowtie_Indexed/Kaki1_
v2.2_concat_index.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-L1_S3_L001.bam

```

### F.2.3.2: Pied Stilt

Available at: [https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/Bowtie2\\_PiedStilt.sh](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/Bowtie2_PiedStilt.sh)

```
#!/bin/sh
```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383_Concat_1set_AusPied.b
am

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384_Concat_1set_AusPied.b
am

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
L1_S43_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
L1_S43_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385_Concat_1set_AusPied.b
am

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386_Concat_1set_AusPied.b
am

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387_Concat_1set_AusPied.b
am

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-
L1_S3_L002_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-
L1_S3_L002_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388_Concat_1set_AusPied.b
am

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-
L1_S4_L003_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-
L1_S4_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >

```

```
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389_Concat_1set_AusPied.b  
am
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x  
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f  
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-  
L1_S5_L003_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-  
L1_S5_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390_Concat_1set_AusPied.b  
am
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x  
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f  
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-  
L1_S6_L004_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-  
L1_S6_L004_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391_Concat_1set_AusPied.b  
am
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x  
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f  
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-  
L1_S7_L005_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-  
L1_S7_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392_Concat_1set_AusPied.b  
am
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x  
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f  
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-  
L1_S8_L005_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-  
L1_S8_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393_Concat_1set_AusPied.b  
am
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x  
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f  
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-  
L1_S9_L005_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-  
L1_S9_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394_Concat_1set_AusPied.b  
am
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x  
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f  
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-  
L1_S10_L007_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-  
L1_S10_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395_Concat_1set_AusPied.b  
am
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x  
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f  
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-  
L1_S11_L007_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
```

```

L1_S11_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-

```

```

L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 4 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406_Concat_1set_AusPied.b
am

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R1_001_paired_trimmed.fq -2

```

```

/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416_AusPied.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417_AusPied.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/AusPiedStilt/Aus1_v2.0_Bowtie/Aus1_v2.0_concat_indexed.f
asta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418_AusPied.bam

```

### F.2.3.3: Pied Avocet

Available at:

[https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/Bowtie2\\_PiedAvocet.sh](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/Bowtie2_PiedAvocet.sh)

```
#!/bin/sh
```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383_Concat_1set_Avocet.ba
m

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384_Concat_1set_Avocet.ba
m

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
L1_S43_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
L1_S43_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385_Concat_1set_Avocet.ba
m

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386_Concat_1set_Avocet.ba
m

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >

```

```
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387_Concat_1set_Avocet.bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-
L1_S3_L002_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-
L1_S3_L002_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388_Concat_1set_Avocet.ba
m
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-
L1_S4_L003_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-
L1_S4_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389_Concat_1set_Avocet.ba
m
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-
L1_S5_L003_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-
L1_S5_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390_Concat_1set_Avocet.ba
m
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-
L1_S6_L004_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-
L1_S6_L004_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391_Concat_1set_Avocet.ba
m
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-
L1_S7_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-
L1_S7_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392_Concat_1set_Avocet.ba
m
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-
L1_S8_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-
L1_S8_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393_Concat_1set_Avocet.ba
m
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-
L1_S9_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-
```



```

L1_S9_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-

```

```

L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406_Concat_1set_Avocet.ba
m

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome

```

```

1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |

```

```

/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417_Avocet.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Avocet_Concat/bowtie/avoset.v1.superscaffolds.chromosome
1_indexed.fasta -1 /share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418_Avocet.bam

```

#### F.2.3.4: Killdeer

Available at:

[https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/Bowtie2\\_Killdeer.sh](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/Bowtie2_Killdeer.sh)

```

#!/bin/sh

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383-
L1_S41_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01383_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384-
L1_S42_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01384_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
L1_S43_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385-
L1_S43_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >

```

```
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01385_Concat_1set_Killdeer.  
bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x  
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-  
L1_S1_L001_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386-  
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01386_Concat_1set_Killdeer.  
bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x  
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-  
L1_S2_L001_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387-  
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01387_Concat_1set_Killdeer.  
bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x  
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-  
L1_S3_L002_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388-  
L1_S3_L002_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01388_Concat_1set_Killdeer.  
bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x  
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-  
L1_S4_L003_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389-  
L1_S4_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01389_Concat_1set_Killdeer.  
bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x  
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-  
L1_S5_L003_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390-  
L1_S5_L003_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01390_Concat_1set_Killdeer.  
bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x  
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-  
L1_S6_L004_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391-  
L1_S6_L004_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |  
/share/data3/conda/miniconda3/bin/samtools view -bS - >  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01391_Concat_1set_Killdeer.  
bam
```

```
/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x  
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-  
L1_S7_L005_R1_001_paired_trimmed.fq -2  
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392-
```

```

L1_S7_L005_R2_001paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01392_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-
L1_S8_L005_R1_001paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393-
L1_S8_L005_R2_001paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01393_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-
L1_S9_L005_R1_001paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394-
L1_S9_L005_R2_001paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01394_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R1_001paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395-
L1_S10_L007_R2_001paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01395_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R1_001paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396-
L1_S11_L007_R2_001paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01396_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R1_001paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397-
L1_S12_L007_R2_001paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01397_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R1_001paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398-
L1_S1_L001_R2_001paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01398_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-

```

```

L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01399_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01400_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01401_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01402_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01403_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404-
L1_S7_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01404_Concat_1set_Killdeer.
bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405-
L1_S8_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01405_Concat_1set_Killdeer.
bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406-
L1_S9_L005_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01406_Concat_1set_Killdeer.
bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407-
L1_S35_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01407_Killdeer.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408-
L1_S36_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01408_Killdeer.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409-
L1_S37_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01409_Killdeer.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410-
L1_S38_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01410_Killdeer.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411-
L1_S39_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01411_Killdeer.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412-
L1_S40_L006_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01412_Killdeer.bam

```

```

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R1_001_paired_trimmed.fq -2

```



```

/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413-
L1_S10_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01413_Killdeer.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414-
L1_S11_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01414_Killdeer.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415-
L1_S12_L007_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01415_Killdeer.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416-
L1_S1_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01416_Killdeer.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417-
L1_S2_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01417_Killdeer.bam

/share/data3/conda/miniconda3/bin/bowtie2 -p 32 -x
/share/data3/sga94/Genomes/Killdeer_Concat/bowtie/killdeer_concat_indexed.fasta -1
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R1_001_paired_trimmed.fq -2
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418-
L1_S3_L001_R2_001_paired_trimmed.fq --very-sensitive -X 600 -I 0 |
/share/data3/conda/miniconda3/bin/samtools view -bS - >
/share/data3/sga94/IKMB_Data/fastq/trimmed/trimmomatic/H01418_Killdeer.bam

```

## F.2.4: Samtools 1.9 Sort

### F.2.4.1: Kaki

Available at:

[https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/Samtools\\_Sort\\_Kaki.sh](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/Samtools_Sort_Kaki.sh)

```
#/bin/sh
```

```
samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H013
83_Concat_1set_kaki.bam'
```





```

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
05_Concat_1set_kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01405_Concat_1set_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
06_Concat_1set_kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01406_Concat_1set_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
07_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01407_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
08_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01408_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
09_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01409_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
10_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01410_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
11_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01411_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
12_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01412_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
13_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01413_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
14_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01414_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
15_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01415_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014

```

```

16_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01416_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
17_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01417_Kaki.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage/H014
18_Kaki.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Kaki/1set_and_lowcoverage_sort
ed/H01418_Kaki.sorted'

```

### F.2.4.2: Pied Stilt

Available at:

[https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/Samtools\\_Sort\\_PiedStilt.sh](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/Samtools_Sort_PiedStilt.sh)

```

#/bin/sh

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01383_Concat_1set_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01383_Concat_1set_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01384_Concat_1set_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01384_Concat_1set_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01385_Concat_1set_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01385_Concat_1set_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01386_Concat_1set_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01386_Concat_1set_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01387_Concat_1set_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01387_Concat_1set_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01388_Concat_1set_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01388_Concat_1set_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01389_Concat_1set_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01389_Concat_1set_AusPied.sorted' \

```





```

01411_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01411_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01412_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01412_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01413_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01413_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01414_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01414_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01415_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01415_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01416_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01416_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01417_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01417_AusPied.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage/H
01418_AusPied.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/AusPied/1set_and_lowcoverage_s
orted/H01418_AusPied.sorted'

```

### F.2.4.3: Pied Avocet

Available at:

[https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/Samtools\\_Sort\\_Avocet.sh](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/Samtools_Sort_Avocet.sh)

```

#!/bin/sh

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1383_Concat_1set_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01383_Concat_1set_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1384_Concat_1set_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01384_Concat_1set_Avocet.sorted' \

```







```

1406_Concat_1set_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01406_Concat_1set_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1407_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01407_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1408_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01408_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1409_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01409_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1410_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01410_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1411_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01411_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1412_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01412_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1413_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01413_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1414_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01414_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1415_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01415_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1416_Avocet.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_so
rted/H01416_Avocet.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H0
1417_Avocet.bam'

```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_sorted/H01417_Avocet.sorted' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage/H01418_Avocet.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Avocet/1set_and_lowcoverage_sorted/H01418_Avocet.sorted'
```

#### F.2.4.4: Killdeer

Available at:

[https://github.com/sgalla32/Resequence\\_Data\\_Processing/blob/master/Samtools\\_Sort\\_Killdeer.sh](https://github.com/sgalla32/Resequence_Data_Processing/blob/master/Samtools_Sort_Killdeer.sh)

```
#!/bin/sh
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01393_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01393_Concat_1set_Killdeer.sorted' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01395_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01395_Concat_1set_Killdeer.sorted.bam' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01396_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01396_Concat_1set_Killdeer.sorted.bam' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01397_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01397_Concat_1set_Killdeer.sorted.bam' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01397_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01397_Concat_1set_Killdeer.sorted.bam' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01398_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01398_Concat_1set_Killdeer.sorted.bam' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01399_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01399_Concat_1set_Killdeer.sorted.bam' \
```

```
samtools sort
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01400_Concat_1set_Killdeer.bam'
```

```
 '/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H01400_Concat_1set_Killdeer.sorted.bam' \
```



```

deer.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H014
12_Killdeer.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01413_Kill
deer.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H014
13_Killdeer.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01414_Kill
deer.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H014
14_Killdeer.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01415_Kill
deer.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H014
15_Killdeer.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01416_Kill
deer.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H014
16_Killdeer.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01417_Kill
deer.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H014
17_Killdeer.sorted' \

samtools sort
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Bam_Files/H01418_Kill
deer.bam'
'/media/stephanie/Data/Resequencing/Bowtie_BAM_Files/Killdeer/Sorted_Bam_Files/H014
18_Killdeer.sorted' \

```

### ***F.2.5: BCFTools v. 1.9 Mpileup***

### ***F.2.6 BCFTools & VCFTools v. 1.9***

Available at:

[https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/BCFTools\\_VCFTools\\_Resequencing.sh](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/BCFTools_VCFTools_Resequencing.sh)

```

#/bin/sh

#Kaki variant calls.

bcftools call Kaki.bcf -mv -Ob -o Kaki_VariantCalls.bcf

bcftools view Kaki_VariantCalls.bcf -v snps -m 2 -M 2 -q 0.05:minor -i
'AVG(FMT/DP)>10 & %QUAL>20' -o
Kaki_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf

bcftools +prune -l 0.8 -w 1000 -e "F_MISSING>=0.1"
Kaki_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf -Ov -o
Kaki_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missin
g0.1.vcf

```

```

vcftools --vcf
Kaki_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1.vcf --thin 150 --out
Kaki_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1_thin150.vcf --recode

#Aus. pied stilt variant calls.

bcftools call AusPied.bcf -mv -Ob -o AusPied_VariantCalls.bcf

bcftools view AusPied_VariantCalls.bcf -v snps -m 2 -M 2 -q 0.05:minor -i
'AVG(FMT/DP)>10 & %QUAL>20' -o
AusPied_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf

bcftools +prune -l 0.8 -w 1000 -e "F_MISSING>=0.1"
AusPied_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf -Ov -o
AusPied_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1.vcf

vcftools --vcf
AusPied_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1.vcf --thin 150 --out
AusPied_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1_thin150.vcf --recode

#Pied avocet variant calls.

bcftools call Avocet.bcf -mv -Ob -o Avocet_VariantCalls.bcf

bcftools view Avocet_VariantCalls.bcf -v snps -m 2 -M 2 -q 0.05:minor -i
'AVG(FMT/DP)>10 & %QUAL>20' -o
Avocet_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf

bcftools +prune -l 0.8 -w 1000 -e "F_MISSING>=0.1"
Avocet_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf -Ov -o
Avocet_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1.vcf

vcftools --vcf
Avocet_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1.vcf --thin 150 --out
Avocet_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1_thin150.vcf --recode

#Killdeer variant calls.

bcftools call Killdeer.bcf -mv -Ob -o Kaki_VariantCalls.bcf

bcftools view Killdeer_VariantCalls.bcf -v snps -m 2 -M 2 -q 0.05:minor -i
'AVG(FMT/DP)>10 & %QUAL>20' -o
Killdeer_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf

bcftools +prune -l 0.8 -w 1000 -e "F_MISSING>=0.1"
Killdeer_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf -Ov -o
Killdeer_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1.vcf

vcftools --vcf
Killdeer_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1.vcf --thin 150 --out
Killdeer_FinalVariantCalls_BCFTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.8_Missing0.1_thin150.vcf --recode

```

## F.3 R-Scripts

The R scripts provided here show the general commands used for Chapter 4.

### F.3.1: KGD - General Script for GBS and Resequencing

KGD can be downloaded from this repository: <https://github.com/AgResearch/KGD>

```
#If you are converting a vcf from Tassel5 or STACKS, remember to convert the vcf to
an ra.tab file using vcf2ra_ro_ao.py
#Use the following command to remove anything from the environment:
rm(list = ls())
#Set your working directory for your file locations.
setwd("FILE_LOCATION")
genofile <- "your.dataset.vcf.ra.tab"
gform <- "Tassel"
source("GBS-Chip-Gmatrix.R")
Gfull <- calcG()
GHWdgm.05 <- calcG(which(HWdis > -0.05), "HWdgm.05", npc = 4)
str(GHWdgm.05)
#Write out your relatedness estimates in vector form.
writeG(Gfull, "Output", outtype = 3)
#Write out your relatedness estimates in matrix form.
writeG(Gfull, "Output", outtype = 2, ,seqID)
```

### F.3.2: Relatedness Scale Transformation & Mantel Tests

These general scripts were used for both GBS and resequencing.

```
#Removes everything from the R environment. Only use when fresh start needed.
rm(list=ls())

#Reads your csv relatedness matrix into R.
M <- as.matrix(read.csv("Relatedness_NoHeaders.csv", header = FALSE))

#The function diag(x) extracts or replaces the diagonal of a matrix, or construct a
diagonal matrix.
#x can specify a matrix, when it extracts the diagonal.
#x can be missing and nrow is specified, it returns an identity matrix
#x can be scalar (length-one vector) and the only argument it returns a square
identity matrix of size given by the scalar.
#x can be a ???numeric??? (complex, numeric, integer, logical, or raw) vector,
either of length at least 2 or there were further arguments. This returns a matrix
with the given diagonal and zero off-diagonal entries.
D <- diag(36) * 1/sqrt(diag(M))

#This creates a scaled matrix with ones as the diagonals.
#Note, % is an infix operator that does math in the background (in this case,
multiplying).
ScaledMatrix <- D %*% M %*% D

#Converts a covariate matrix to a correlation matrix.
ScaledMatrix_Correlation <- cov2cor(ScaledMatrix)

write.table(ScaledMatrix_Correlation, file="ScaledMatrix.txt", sep= " ")

write.csv(ScaledMatrix_Correlation, file="ScaledMatrix.csv")

#####
#Mantel Test
#####
```



```

#Make sure the programme Ape is installed.
M1 <- as.matrix(read.csv("Kaki_Scaled_Matrix_Correlation_NoIDs.csv", header =
FALSE))
M2 <- as.matrix(read.csv("AusPied_Scaled_Matrix_Correlation_NoIDs.csv", header =
FALSE))
M3 <- as.matrix(read.csv("Avocet_Scaled_Matrix_Correlation_NoIDs.csv", header =
FALSE))
M4 <- as.matrix(read.csv("Killdeer_Scaled_Matrix_Correlation_NoIDs.csv", header =
FALSE))

sink("MantelTests.txt")
mantel.test(M1, M2, nperm = 1000, graph = TRUE,
            alternative = "two.sided")
mantel.test(M1, M3, nperm = 1000, graph = TRUE,
            alternative = "two.sided")
mantel.test(M1, M4, nperm = 1000, graph = TRUE,
            alternative = "two.sided")
sink()

```

### *F.3.3: Descriptive Statistic Comparisons*

These general scripts were used for GBS and resequencing.

```

#Removes everything from the environment
rm(list=ls())

#Check packages for dplyr, pastecs, ggplot2, datasets, graphics, stats, ggpubr

#####
#Pi_Data
#####
#Create a dataset (e.g., "Pi_Data")
Pi_Data <- read.csv("Pi_2Columns.csv", sep = ",", header = TRUE)

# Show a random sample to check your data.
set.seed(1234)
dplyr::sample_n(Pi_Data, 10)

#Allows you to rename your first two column headers, if you need to.
names(Pi_Data) <- c("Reference", "Pi")

#Lets you check to make sure your columns are correctly labeled.
head(Pi_Data)

# Show the levels (or variables that you have)
levels(Pi_Data$Reference)

#If the levels are not automatically in the correct order, re-order them as
follows:
Pi_Data$Reference <- ordered(Pi_Data$Reference,
                           levels = c("Kaki", "AusPied", "Avocet", "Killdeer"))

#Creates a box plot
ggplot(Pi_Data, aes(x = Reference, y = Pi)) + geom_boxplot()

Pi_Data_Desc <- read.csv("Pi.csv", sep = ",", header = TRUE)

#Sink allows you to write your analyses to an output file.
sink('Pi_DescriptiveStats.txt')

#Get min, 1st quartile, median, mean, 3rd quartile, max.
summary(Pi_Data_Desc)

#Make sure the package pastecs is clicked. The following line will give you all
sorts of descriptive stats, and will output it to a new dataset.

```

```

stat.desc(Pi_Data_Desc, basic = TRUE, desc = TRUE, p=0.95)

#Finish the output file here.
sink()

#####
#Het_Data
#####

#Create a dataset (e.g., "Het_Data").
Het_Data <- read.csv("Het_2Columns.csv", sep = ",", header = TRUE)

# Show a random sample to check your data.
set.seed(1234)
dplyr::sample_n(Het_Data, 10)

#Allows you to rename your first two column headers, if you need to.
names(Het_Data) <- c("Reference", "Het")

#Lets you check to make sure your columns are correctly labeled.
head(Het_Data)

# Show the levels (or variables that you have)
levels(Het_Data$Reference)

#If the levels are not automatically in the correct order, re-order them as
follows:
Het_Data$Reference <- ordered(Het_Data$Reference,
                             levels = c("Kaki", "AusPied", "Avocet", "Killdeer"))

#Creates a box plot
ggplot(Het_Data, aes(x = Reference, y = Het)) + geom_boxplot()

#Create a dataset (e.g., "MyData") which has your text file, delineated as a tab-
separated file with no headers.
Het_Data_Desc <- read.csv("Het.csv", sep = ",", header = TRUE)

#Lets you check to make sure your columns are correctly labeled.
head(Het_Data_Desc)

#Sink allows you to write your analyses to an output file.
sink('Het_DescriptiveStats.txt')

#Get min, 1st quartile, median, mean, 3rd quartile, max.
summary(Het_Data_Desc)

#Make sure the package pastecs is clicked. The following line will give you all
sorts of descriptive stats, and will output it to a new dataset.
stat.desc(Het_Data_Desc, basic = TRUE, desc = TRUE, norm = TRUE, p=0.95)

#Finish the output file here.
sink()

#####
#Relatedness_Data
#####

#Create a dataset (e.g., "MyData") which has your text file, delineated as a tab-
separated file with no headers.
Relatedness_Data <- read.csv("R_2Columns.csv", sep = ",", header = TRUE)

# Show a random sample to check your data.
set.seed(1234)
dplyr::sample_n(Relatedness_Data, 10)

#Allows you to rename your first two column headers, if you need to.
names(Relatedness_Data) <- c("Reference", "R")

```

```

#Lets you check to make sure your columns are correctly labeled.
head(Relatedness_Data)

# Show the levels (or variables that you have)
levels(Relatedness_Data)

#If the levels are not automatically in the correct order, re-order them as follow:
Relatedness_Data$Reference <- ordered(Relatedness_Data$Reference,
                                     levels = c("Kaki", "AusPied", "Avocet", "Killdeer"))

#Creates a box plot
ggplot(Relatedness_Data, aes(x = Reference, y = R)) + geom_boxplot()

#Create a dataset (e.g., "MyData") which has your text file, delineated as a tab-
separated file with no headers.
Relatedness_Data_Desc <- read.csv("R.csv", sep = ",", header = TRUE)

#Sink allows you to write your analyses to an output file.
sink('Relatedness_Data_Desc_Scaled.txt')

#Get min, 1st quartile, median, mean, 3rd quartile, max.
summary(Relatedness_Data_Desc)

#Make sure the package pastecs is clicked. The following line will give you all
sorts of descriptive stats, and will output it to a new dataset.
stat.desc(Relatedness_Data_Desc, basic = TRUE, desc = TRUE, norm = TRUE, p=0.95)

#Finish the output file here.
sink()

#####
#Depth
#####

#Create a dataset (e.g., "MyData") which has your text file, delineated as a tab-
separated file with no headers.
Depth <- read.csv("Depth.csv", sep = ",", header = TRUE)

#Sink allows you to write your analyses to an output file.
sink('Depth_DescStats.txt')

#Get min, 1st quartile, median, mean, 3rd quartile, max.
summary(Depth)

#Make sure the package pastecs is clicked. The following line will give you all
sorts of descriptive stats, and will output it to a new dataset.
stat.desc(Depth, basic = TRUE, desc = TRUE, norm = TRUE, p=0.95)

#Finish the output file here.
sink()

#####
#Missingness
#####

#Create a dataset (e.g., "MyData") which has your text file, delineated as a tab-
separated file with no headers.
Missing <- read.csv("Missing.csv", sep = ",", header = TRUE)

#Sink allows you to write your analyses to an output file.
sink('Missing_DescStats.txt')

#Get min, 1st quartile, median, mean, 3rd quartile, max.
summary(Missing)

#Make sure the package pastecs is clicked. The following line will give you all
sorts of descriptive stats, and will output it to a new dataset.
stat.desc(Missing, basic = TRUE, desc = TRUE, norm = TRUE, p=0.95)

```

```
#Finish the output file here.
sink()
```

### ***F.3.4: Correlations***

These general scripts were used for GBS and resequencing.

```
#Make sure that ggpubr is selected.
library("ggpubr")

#Remove anything that is in the environment
rm(list=ls())

#####
#Heterozygosity/Inbreeding
#####

#Create a data set called Het_Data for individual heterozygosity/inbreeding.
Het_Data <- read.csv("Het.csv", sep = ",", header = TRUE)

#Check to make sure that Het_Data uploaded correctly.
head(Het_Data, 6)

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we will
use the ggpubr R package.
#Adds regression line of best fit and 95% confidence interval
ggscatter(Het_Data, x = "Kaki", y = "AusPied",
          add = "reg.line", conf.int = TRUE,
          xlab = "Kaki", ylab = "AusPied")

ggscatter(Het_Data, x = "Kaki", y = "Avocet",
          add = "reg.line", conf.int = TRUE,
          xlab = "Kaki", ylab = "Avocet")

ggscatter(Het_Data, x = "Kaki", y = "Killdeer",
          add = "reg.line", conf.int = TRUE,
          xlab = "Kaki", ylab = "Killdeer")

#Is the covariation linear? Yes, form the plot above, the relationship is linear.
#In the situation where the scatter plots show curved patterns, we are dealing with
nonlinear association between the two variables.
#Are the data from each of the 2 variables (x, y) follow a normal distribution?

sink('shapiro-wilk_normality_hetdata.txt')
# Shapiro-Wilk normality test for Kaki
shapiro.test(Het_Data$Kaki)
# Shapiro-Wilk normality test for PiedStilt
shapiro.test(Het_Data$AusPied)
# Shapiro-Wilk normality test for Avocet
shapiro.test(Het_Data$Avocet)
# Shapiro-Wilk normality test for Killdeer
shapiro.test(Het_Data$Killdeer)

sink()

#Visual inspection of the data normality using Q-Q plots (quantile-quantile
plots).
#Q-Q plot draws the correlation between a given sample and the normal distribution.

# Kaki
ggqqplot(Het_Data$Kaki, ylab = "Kaki")
# Kaki
ggqqplot(Het_Data$PiedStilt, ylab = "AusPied")
# Kaki
```

```

ggqqplot(Het_Data$Avocet, ylab = "Avocet")
# Kaki
ggqqplot(Het_Data$Killdeer, ylab = "Killdeer")

#Pearson's Correlation test between Het and Reference variables:
sink('Correlation_hetdata.txt')

Kaki_v_PiedStilt_Pearsons <- cor.test(Het_Data$Kaki, Het_Data$AusPied,
                                     method = "pearson")
Kaki_v_PiedStilt_Pearsons

Kaki_v_Avocet_Pearsons <- cor.test(Het_Data$Kaki, Het_Data$Avocet,
                                  method = "pearson")
Kaki_v_Avocet_Pearsons

Kaki_v_Killdeer_Pearsons <- cor.test(Het_Data$Kaki, Het_Data$Killdeer,
                                    method = "pearson")
Kaki_v_Killdeer_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Kaki_v_PiedStilt_Kendall <- cor.test(Het_Data$Kaki, Het_Data$AusPied,
                                    method = "kendall")
Kaki_v_PiedStilt_Kendall

Kaki_v_Avocet_Kendall <- cor.test(Het_Data$Kaki, Het_Data$Avocet,
                                 method = "kendall")
Kaki_v_Avocet_Kendall

Kaki_v_Killdeer_Kendall <- cor.test(Het_Data$Kaki, Het_Data$Killdeer,
                                   method = "kendall")
Kaki_v_Killdeer_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Kaki_v_PiedStilt_Spearman <- cor.test(Het_Data$Kaki, Het_Data$AusPied,
                                     method = "spearman")
Kaki_v_PiedStilt_Spearman

Kaki_v_Avocet_Spearman <- cor.test(Het_Data$Kaki, Het_Data$Avocet,
                                  method = "spearman")
Kaki_v_Avocet_Spearman

Kaki_v_Killdeer_Spearman <- cor.test(Het_Data$Kaki, Het_Data$Killdeer,
                                    method = "spearman")
Kaki_v_Killdeer_Spearman

sink()

#####
#Relatedness
#####

#Create a dataset called R_Data.
R_Data <- read.csv("R.csv", sep = ",", header = TRUE)

#Check to make sure R_Data uploaded correctly.
head(R_Data, 6)

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll
use the ggpubr R package.
ggscatter(R_Data, x = "Kaki", y = "AusPied",
          add = "reg.line", conf.int = TRUE,
          xlab = "Kaki", ylab = "AusPied")

ggscatter(R_Data, x = "Kaki", y = "Avocet",
          add = "reg.line", conf.int = TRUE,
          xlab = "Kaki", ylab = "Avocet")

ggscatter(R_Data, x = "Kaki", y = "Killdeer",

```

```

        add = "reg.line", conf.int = TRUE,
        xlab = "Kaki", ylab = "Killdeer")

#Is the covariation linear? Yes, form the plot above, the relationship is linear.
#In the situation where the scatter plots show curved patterns, we are dealing with
nonlinear association between the two variables.
#Are the data from each of the 2 variables (x, y) follow a normal distribution?

sink("relatedness_shapiro.test.txt")
# Shapiro-Wilk normality test for Kaki
shapiro.test(R_Data$Kaki)
# Shapiro-Wilk normality test for PiedStilt
shapiro.test(R_Data$AusPied)
# Shapiro-Wilk normality test for Avocet
shapiro.test(R_Data$Avocet)
# Shapiro-Wilk normality test for Killdeer
shapiro.test(R_Data$Killdeer)
sink()

#Visual inspection of the data normality using Q-Q plots (quantile-quantile
plots).
#Q-Q plot draws the correlation between a given sample and the normal distribution.

# Kaki
ggqqplot(R_Data$Kaki, ylab = "Kaki")
# Kaki
ggqqplot(R_Data$PiedStilt, ylab = "PiedStilt")
# Kaki
ggqqplot(R_Data$Avocet, ylab = "Avocet")
# Kaki
ggqqplot(R_Data$Killdeer, ylab = "Killdeer")

sink("Relatedness_Correlations_Scaled.txt")

#Pearson's Correlation test between R and Reference variables:
Kaki_v_PiedStilt_Pearsons <- cor.test(R_Data$Kaki, R_Data$AusPied,
                                     method = "pearson")
Kaki_v_PiedStilt_Pearsons

Kaki_v_Avocet_Pearsons <- cor.test(R_Data$Kaki, R_Data$Avocet,
                                  method = "pearson")
Kaki_v_Avocet_Pearsons

Kaki_v_Killdeer_Pearsons <- cor.test(R_Data$Kaki, R_Data$Killdeer,
                                    method = "pearson")
Kaki_v_Killdeer_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Kaki_v_PiedStilt_Kendall <- cor.test(R_Data$Kaki, R_Data$AusPied,
                                    method = "kendall")
Kaki_v_PiedStilt_Kendall

Kaki_v_Avocet_Kendall <- cor.test(R_Data$Kaki, R_Data$Avocet,
                                  method = "kendall")
Kaki_v_Avocet_Kendall

Kaki_v_Killdeer_Kendall <- cor.test(R_Data$Kaki, R_Data$Killdeer,
                                   method = "kendall")
Kaki_v_Killdeer_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Kaki_v_PiedStilt_Spearman <- cor.test(R_Data$Kaki, R_Data$AusPied,
                                      method = "spearman")
Kaki_v_PiedStilt_Spearman

Kaki_v_Avocet_Spearman <- cor.test(R_Data$Kaki, R_Data$Avocet,
                                   method = "spearman")
Kaki_v_Avocet_Spearman

```

```
Kaki_v_Killdeer_Spearman <- cor.test(R_Data$Kaki, R_Data$Killdeer,
                                     method = "spearman")
Kaki_v_Killdeer_Spearman

sink()
```

### *F.3.5: Heterozygosity Linear Mixed Effects Model*

```
# Mixed Effects Models in R

#Removes everything from the environment. Only use when fresh start needed.
rm(list=ls())

library(lme4)
library(lmerTest)

#Set your working directory
setwd("LOCATION")

# If .csv tab file, use this:
HeterozygosityData <- read.csv("Het_Lme4.csv", sep = ",", header = TRUE)

#Check to make sure your labels are correct.
head(HeterozygosityData, 6)

#Identify the structure of the R dataset.
str(HeterozygosityData)

#See which variables belong to Reference.
levels(HeterozygosityData$Reference)

#Set the order of variables.
HeterozygosityData$Reference <- factor(HeterozygosityData$Reference,
                                       levels = c("Kaki", "AusPied", "Avocet", "Killdeer"))

library(ggplot2)

ggplot(HeterozygosityData, aes(x=Reference, y=Het)) +
  geom_boxplot()
  ylab("Het") + xlab("Reference") +
  theme_bw()

#####
# Random intercept model
#####
#The 1 before reference defines that the first reference is the intercept.
#It also delineates it on a population level.
hetmod1 <- lmer(Het ~ 1 + Reference + MeanKinship + (1|Individual),
               data=HeterozygosityData)

#Note that sink writes all your results to a file, and closes the file at the end.

sink('hetmod1summary_anova_tukey.txt')
summary(hetmod1)
anova(hetmod1)
sink()

sink('hetmod2summary_anova_tukey.txt')
hetmod2 <- lmer(Het ~ 1 + Reference + (1|Individual), data=HeterozygosityData)
summary(hetmod2)
anova(hetmod2)
sink()

#How to do a multiple comparison (Tukey) using multcomp.

library(multcomp)
```

```

sink('hetmod1summary_anova_tukey.txt', append = TRUE)
summary(glht(hetmod1, linfct = mcp(Reference="Tukey")))
sink()

sink('hetmod2summary_anova_tukey.txt', append = TRUE)
summary(glht(hetmod2, linfct = mcp(Reference="Tukey")))
sink()

library(lattice)
#residuals vs. fitted. The main thing to look for here is to look for a starry
night. It should look random.
#The one here is a bit clustered in the middle.
plot(hetmod1, results="hide", fig.show='hide')

#residuals vs. fitted by Individual
#Take a look at your four points for each individual.
plot(hetmod1, resid(., scaled=TRUE) ~ fitted(.) | Individual, abline = 0)

# QQplot for residuals
qqmath(hetmod1)

#QQplot for the random intercepts
#You want all the dots to fall along the line. This is how you know the model
fits.
qqmath(ranef(hetmod1, whichel="Individual"))
#Which they kind of do! The tails fall off a bit, so we might have to transform the
data.

```



## Appendix G: Chapter 5, Supplemental Materials

Supplemental information for: A comparison of pedigree, genetic, and genomic estimates of relatedness for informing pairing decisions in two critically endangered birds: Implications for conservation breeding programmes worldwide.

Authors: Stephanie Galla<sup>1\*</sup>, Roger Moraga<sup>2</sup>, Liz Brown<sup>3</sup>, Simone Cleland<sup>3</sup>, Marc P. Hoepfner<sup>4</sup>, Richard F. Maloney<sup>5</sup>, Anne Richardson<sup>6</sup>, Lyndon Slater<sup>7</sup>, Anna W. Santure<sup>8</sup>, Tammy E. Steeves<sup>1</sup>

<sup>1</sup> School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

<sup>2</sup> Tea Break Bioinformatics, Ltd., Palmerston North, New Zealand

<sup>3</sup> New Zealand Department of Conservation, Twizel, New Zealand

<sup>4</sup> Institute for Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany

<sup>5</sup> New Zealand Department of Conservation, Christchurch, New Zealand

<sup>6</sup> The Isaac Conservation and Wildlife Trust, Christchurch, New Zealand

<sup>7</sup> New Zealand Department of Conservation, Rangiora, New Zealand

<sup>8</sup> School of Biological Sciences, University of Auckland, Auckland, New Zealand

\* Corresponding author, [sgalla32@gmail.com](mailto:sgalla32@gmail.com)

### G.1: COANCESTRY Microsatellite Simulations

The programme COANCESTRY v. 1.0.1.9 (Wang et al. 2011) offers seven different estimators of relatedness for genetic and genomic markers, and to choose the most appropriate estimator for the kakī and kākārīki karaka microsatellite datasets, we employed the simulation module within COANCESTRY using allele frequencies, missing data, and error rates from our microsatellite datasets. To produce dyads that represent the relationships and degree of inbreeding found within kakī and kākārīki karaka, we used the R package ‘identity’ (Li 2010) to generate 10,879 dyads for kakī and 1,484 dyads for kākārīki karaka based on the pedigrees of both species. The frequency of each unique dyad in the kakī and kākārīki karaka datasets were scaled to create 1,000 dyads for each set that are representative of relationships between individuals used in this study. The COANCESTRY simulations were conducted using allele frequencies, error rates, and missing data rates from each microsatellite data set, with settings changed to account for inbreeding. The

triadic likelihood approach (Wang 2007) was selected given it had the highest Pearson's correlation with 'true' relatedness and lowest variance for both kakī (Table G1) and kākāriki karaka (Table G2), as per Hammerly et al. (2013).

Table G1. Average relatedness, variance, and correlation to 'true' relatedness for seven relatedness estimators using simulated datasets from kakī microsatellite allele frequencies and dyads in the kakī pedigree.

	TrioEst	WEst	LLEst	LREst	REst	QGEst	MEst	Actual Relatedness
Average	0.214	0.063	0.064	0.067	0.065	0.064	0.236	0.07
Variance	<b>0.047</b>	0.131	0.125	0.088	0.098	0.103	0.051	0.023
Pearson's R	<b>0.511</b>	0.407	0.418	0.483	0.473	0.451	0.502	1

Table G2. Average relatedness, variance, and correlation to 'true' relatedness for seven relatedness estimators using simulated datasets from kākāriki karaka microsatellite allele frequencies and dyads in the kākāriki karaka pedigree.

	TrioEst	WEst	LLEst	LREst	REst	QGEst	MEst	Actual Relatedness
Average	0.283	0.145	0.149	0.159	0.15	0.152	0.315	0.151
Variance	<b>0.07</b>	0.168	0.166	0.123	0.158	0.146	0.076	0.051
Pearson's R	<b>0.675</b>	0.567	0.568	0.663	0.533	0.597	0.661	1

## G.2: Compute Specifications for Genome Assembly

To assemble the kākāriki karaka reference genome, the following compute specifications were used: a 24 central processing unit (CPU) AMD Threadripper control unit (CU), 128 Gb RAM, 2.7 Tb Solid-State hard disk space were used. The run time for the assemblies was 30 hours for Meraculous, and about 121 hours for Masurca, per iteration. Several iterations were performed to compare assemblies with different parameters.

## G.3: SNP Filtering

Base filtering was applied to all tried filtering datasets, including a filtering to retain only biallelic SNPs with a minor allele frequency (MAF) greater than 0.05, a quality score greater than 20, and maximum missingness of 10% per site. This base filtering was designed to increase the quality, completeness, and reliability of SNPs by removing low quality markers and potential artefacts captured by rare multiallelic sites (Campbell et al. 2016) or low frequency sequencing error.

Different depth filters for each site were tested for each dataset to achieve an average of ~10x depth across all sites for each individual, to provide more certainty around reference-guided SNPs used here. Preliminary testing with kakī revealed that using an average minimum depth of 10x was sufficient for meeting this criteria (e.g., the lowest per individual average depth observed post filtering was 9.6x, with mean depth across all sites and individuals of  $28.7x \pm 10.29$  SD). This depth was also used for an aligned study (see Galla et al. 2019). Applying this same filtering scheme to the kākāriki karaka dataset did not achieve an average depth of ~10x across all sites for each individual (the minimum average depth = 5.44x, mean depth across all sites and individuals =  $15.93x \pm 10.11$ ). Therefore, a filtering trial using two different depth filtering strategies was employed, with one filtering using a hard minimum cut-off of 5x depth and the other using an average minimum depth of 20x. Both of these filtering trials also employed a hard maximum cutoff of 200x depth, to filter obvious high-coverage sites that come from collapsed repeats in the reference genome. Because there are known parent-offspring relationships represented in our dataset, and parent-offspring genomic contribution is 50%, we used relatedness between parents and offspring as a biologically meaningful measure to understand which approach best approximated 0.5 with the greatest precision. It was found that a hard minimum cut-off of 5x per SNP resulted in relatedness estimates that were more accurate and precise (Table G3).

Table G3: Results from filtering trials for depth,  $r^2$ , and HWE using the kākāriki karaka data set. Base filtering refers to filter steps for biallelic SNPs with a MAF of 0.05, and missingness of 0.1. Scaled refers to the chosen data set, after scaling for self-relatedness to be equal to 1. Bold text refers to filtering scheme chosen. Alternating blue and white cell fill denotes trials with aligned depth and  $r^2$  settings.

Trial	Average $R \pm SD$	Average PO $R \pm SD$	Min PO $R$	Max PO $R$
Base filtering, minimum depth of 5x, maximum depth 200x, $r^2$ 0.4	$0.25 \pm 0.1$	$0.43 \pm 0.04$	0.36	0.57
Base filtering, minimum depth of 5x, maximum depth 200x, $r^2$ 0.4, HWE	$0.02 \pm 0.22$	$0.43 \pm 0.11$	0.19	0.67
<b>Base filtering, minimum depth of 5x, maximum depth 200x, <math>r^2</math> 0.6</b>	<b><math>0.24 \pm 0.10</math></b>	<b><math>0.43 \pm 0.04</math></b>	<b>0.35</b>	<b>0.58</b>
<b>Base filtering, minimum depth of 5x, maximum depth 200x, <math>r^2</math> 0.6, Scaled</b>	<b><math>0.29 \pm 0.12</math></b>	<b><math>0.53 \pm 0.03</math></b>	<b>0.47</b>	<b>0.67</b>
<b>Base filtering, minimum depth of 5x, maximum depth 200x, <math>r^2</math> 0.6, HWE</b>	<b><math>0.02 \pm 0.22</math></b>	<b><math>0.43 \pm 0.12</math></b>	<b>0.19</b>	<b>0.68</b>
Base filtering, minimum depth of 5x, maximum depth 200x, $r^2$ 0.8	$0.24 \pm 0.11$	$0.44 \pm 0.04$	0.36	0.59
Base filtering, minimum depth of 5x, maximum depth 200x, $r^2$ 0.8, HWE	$0.02 \pm 0.23$	$0.45 \pm 0.12$	0.21	0.7
Base filtering, average minimum depth 20x, maximum depth 200x $r^2$ 0.4	$0.08 \pm 0.13$	$0.33 \pm 0.08$	0.17	0.53
Base filtering, average minimum depth 20x, maximum depth 200x $r^2$ 0.4, HWE	$0.01 \pm 0.19$	$0.36 \pm 0.11$	0.12	0.58
Base filtering, average minimum depth 20x, maximum depth 200x $r^2$ 0.6	$0.05 \pm 0.15$	$0.33 \pm 0.09$	0.14	0.55
Base filtering, average minimum depth 20x, maximum depth 200x $r^2$ 0.6, HWE	$0.01 \pm 0.19$	$0.36 \pm 0.12$	0.12	0.59
Base filtering, average minimum depth 20x, maximum depth 200x $r^2$ 0.8	$0.03 \pm 0.17$	$0.34 \pm 0.07$	0.21	0.57
Base filtering, average minimum depth 20x, maximum depth 200x $r^2$ 0.8, HWE	$0.00 \pm 0.20$	$0.37 \pm 0.12$	0.12	0.62

In addition to depth for kākāriki karaka, different  $r^2$  filters (i.e.,  $r^2 = 0.4, 0.6$ , and  $0.8$ ) for linkage disequilibrium were applied to see how this variable affected relatedness estimates. Further, HWE filters of 0.05 using KGD were applied to each of these trials to see if using a HWE filter on each of these approaches affected the accuracy and precision of

relatedness (Tables G4 and G5). Overall, using a HWE filter resulted in less accurate and precise estimates of relatedness, which may be attributed to our datasets consisting mostly of family groups, thereby violating the assumption of random breeding. Strong to moderate LD filters ( $r^2 = 0.4, 0.6$ ) produced more accurate and precise estimates of relatedness in kakī, but did not make a substantial difference in kākāriki karaka. A moderate LD filter ( $r^2 = 0.6$ ) was chosen for both datasets (Tables G3 and G4).

Table G4: Results from filtering trials for  $r^2$  and HWE using the kakī data set. Base filtering refers to set filters for biallelic SNPs with a MAF of 0.05, an average mean depth of 10x, and missingness of 0.1. Scaled refers to the chosen data set, after scaling for self-relatedness to be equal to 1.

Trial	Average $R$ $\pm$ SD	Average PO $R \pm$ SD	Min PO $R$	Max PO $R$
Base filter, $r^2$ 0.4	0.33 $\pm$ 0.07	0.52 $\pm$ 0.03	0.47	0.60
Base filter, $r^2$ 0.4, HWE	0.02 $\pm$ 0.12	0.37 $\pm$ 0.06	0.30	0.55
<b>Base filter, <math>r^2</math> 0.6</b>	<b>0.25 <math>\pm</math> 0.08</b>	<b>0.47 <math>\pm</math> 0.04</b>	<b>0.42</b>	<b>0.59</b>
<b>Base filter, <math>r^2</math> 0.6, Scaled</b>	<b>0.27 <math>\pm</math> 0.09</b>	<b>0.54 <math>\pm</math> 0.03</b>	<b>0.49</b>	<b>0.61</b>
Base filter, $r^2$ 0.6, HWE	0.01 $\pm$ 0.13	0.38 $\pm$ 0.08	0.28	0.59
Base filter, $r^2$ 0.8	0.16 $\pm$ 0.09	0.42 $\pm$ 0.06	0.34	0.58
Base filter, $r^2$ 0.8, HWE	0.00 $\pm$ 0.13	0.37 $\pm$ 0.09	0.26	0.60

## G.4: SNP-based Relatedness Estimates

To produce pairwise estimates of relatedness using whole-genome SNPs, we used the R script KGD (Dodds et al. 2015), as it was designed to estimate relatedness using reduced-representation and resequence data while taking into account read depth. We also scaled our KGD relatedness values so that self-relatedness was equal to one, for two reasons: 1) Creating a diagonal with a value of 1 simplified Mantel tests performed in relatedness comparison analyses, and 2) Scaling KGD values created parent-offspring relatedness values that approximated 0.5 closer than unscaled KGD relatedness values (Tables S6). Because the inbreeding coefficient ( $F$ ) can be derived from marker-based self relatedness ( $R_{self}$ ) where  $F = R_{self} - 1$  (Dodds et al. 2015), we anticipate that scaling may have

accounted for variance in inbreeding values amongst sampled individuals. This scaling has shown to have minimal bias, as scaled and unscaled KGD relatedness values correlate with one another (Pearson's  $r = 0.99$ ,  $p < 0.001$  for both kakī and kākāriki karaka, Table G7 and G8). Further, downstream MSI scores and MK ranks using KGD scaled and unscaled relatedness values are highly concordant with one another, indicating low bias from scaling (Figure G2).

In order to evaluate the performance of scaled KGD values, estimates of relatedness were compared to other marker-based relatedness estimators, including the triadic likelihood method (i.e., TrioML; Wang 2007), the  $r_{xy}$  estimator (Hedrick & Lacy 2015), and the KING estimator (Waples *et al.* 2019). The TrioML and  $r_{xy}$  methods are particularly applicable to our study systems, as they account for inbreeding in their relatedness estimate. The R package *related* (Pew *et al.* 2015) was used to produce TrioML. Settings were set to account for inbreeding and calculate 95% confidence intervals with a bootstrap value of 100. The  $r_{xy}$  and KING estimators were produced using the programme ngsRelateV2 and final VCFs for resequencing datasets (Hanghøj *et al.* 2019). Known parent/offspring dyads were used as a benchmark to evaluate precision for each approach, as parents/offspring relatedness should approximate 0.5 (Speed & Balding 2015). Results indicate that the scaled KGD approach was able to produce estimates of parent-offspring relatedness that more closely approximated 0.5 than TrioML,  $r_{xy}$ , and KING estimators (Table G3). While KGD estimates were more precise for parent-offspring relationships than these approaches, these estimates still significantly correlate with other estimators for both kakī (Pearson's  $r = 0.87-0.96$ ,  $p < 0.001$ ) and kākāriki karaka (Pearson's  $r = 0.91-0.96$ ,  $p < 0.001$ ; see Tables G7 and G8).

Table G5: SNP-based estimates of relatedness using KGD and TrioML approaches in kakī and kākāriki karaka.

Species	Estimator	Average $R \pm SD$	Min. $R$	Max. $R$	Average Parent-Offspring $R \pm SD$	Min. Parent-Offspring $R$	Max. Parent-Offspring $R$
Kakī	KGD	$0.25 \pm 0.08$	0.11	0.59	$0.48 \pm 0.04$	0.42	0.59
	<b>KGD – Scaled</b>	<b><math>0.27 \pm 0.09</math></b>	<b>0.13</b>	<b>0.61</b>	<b><math>0.54 \pm 0.03</math></b>	<b>0.49</b>	<b>0.61</b>
	TrioML	$0.06 \pm 0.06$	0	0.40	$0.27 \pm 0.06$	0.17	0.40
	KING	$0.18 \pm 0.04$	0.06	0.32	$0.29 \pm 0.01$	0.28	0.32
	$r_{xy}$	$0.09 \pm 0.08$	0	0.5	$0.34 \pm 0.07$	0.23	0.50
Kākāriki karaka	KGD	$0.24 \pm 0.10$	0.07	0.58	$0.43 \pm 0.03$	0.35	0.58
	<b>KGD – Scaled</b>	<b><math>0.29 \pm 0.12</math></b>	<b>0.08</b>	<b>0.67</b>	<b><math>0.53 \pm 0.03</math></b>	<b>0.47</b>	<b>0.67</b>
	TrioML	$0.07 \pm 0.10$	0	0.43	$0.26 \pm 0.04$	0.18	0.43
	KING	$0.16 \pm 0.07$	-0.06	0.33	$0.28 \pm 0.02$	0.22	0.33
	$r_{xy}$	$0.09 \pm 0.12$	0	0.58	$0.34 \pm 0.05$	0.25	0.54

Table G6: Pearson's correlation coefficient between different SNP-based estimates of relatedness in kakī.

	KGD	KGD Scaled	TrioML	Rab	KING
KGD	1	—	—	—	—
KGD Scaled	<b>0.99</b>	1	—	—	—
TrioML	0.96	<b>0.96</b>	1	—	—
$r_{xy}$	0.95	<b>0.96</b>	0.99	1	—
KING	0.80	<b>0.87</b>	0.85	0.86	1

Table S7: Pearson's correlation coefficient between different SNP-based estimates of relatedness in kākāriki karaka.

	KGD	KGD Scaled	TrioML	Rab	KING
KGD	1	—	—	—	—
KGD Scaled	<b>0.99</b>	1	—	—	—
TrioML	0.96	<b>0.96</b>	1	—	—
$r_{xy}$	0.96	<b>0.96</b>	0.99	1	—
KING	0.89	<b>0.91</b>	0.88	0.88	1

## G.5: MSI and MK Rank Correlations

Pearson's correlations between pedigree-, microsatellite-, and SNP-based MSI scores (Figure G1) and MK ranks (Figure G2) were performed in the manuscript. Scatterplots showing these relationships are provided below.

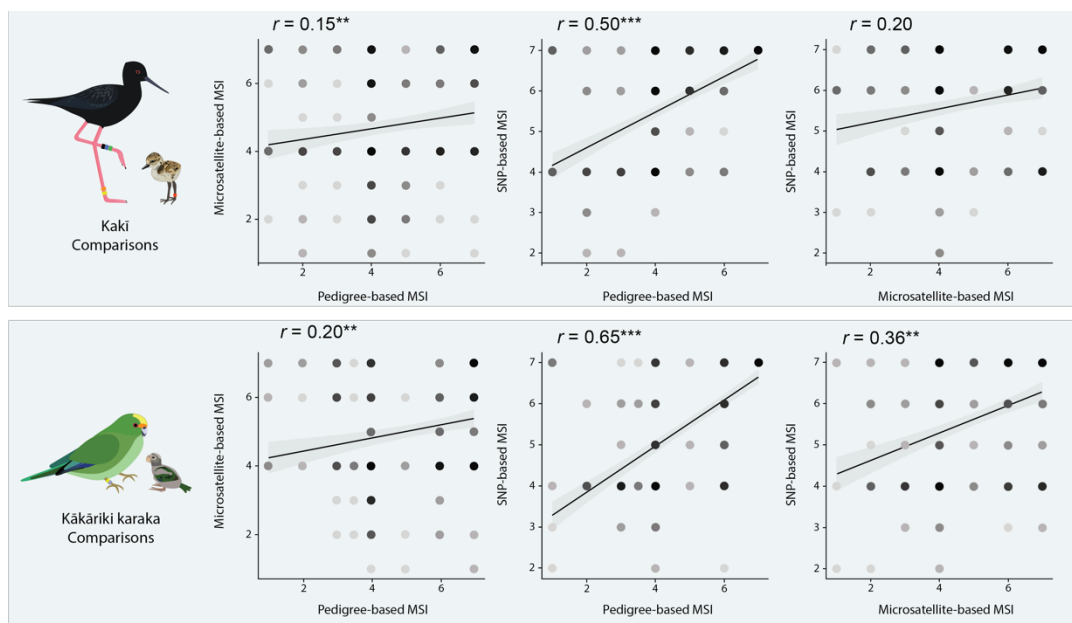


Figure G 1. Scatterplots showing relationships between pedigree-, microsatellite-, and SNP-based MSI scores in kakī and kākārīki karaka. Darker points denote higher frequencies than lighter points. A trend line (black) and 95% confidence intervals (grey) are shown in each comparison.

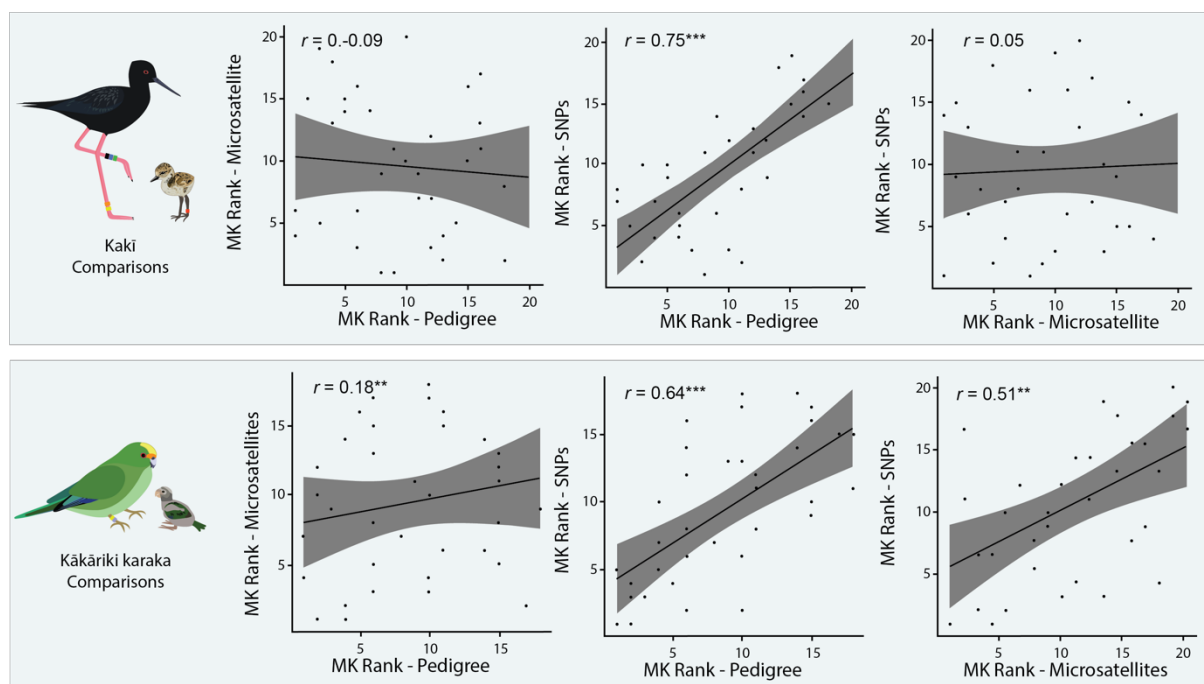


Figure G 2. Scatterplots showing relationships between pedigree-, microsatellite-, and SNP-based MK ranks in kakī and kākārīki karaka. Darker points denote higher frequencies than lighter points. A trend line (black) and 95% confidence intervals (grey) are shown in each comparison. Correlation coefficients are provided in the upper left hand corner of each graph, with \*\* indicating  $p < 0.01$  and \*\*\* indicating  $p < 0.001$ .



Pearson's correlations were also used in the manuscript to understand the downstream effects of scaling on MSI scores and MK rank. Scatterplots showing these relationships are provided below (Figure G3).

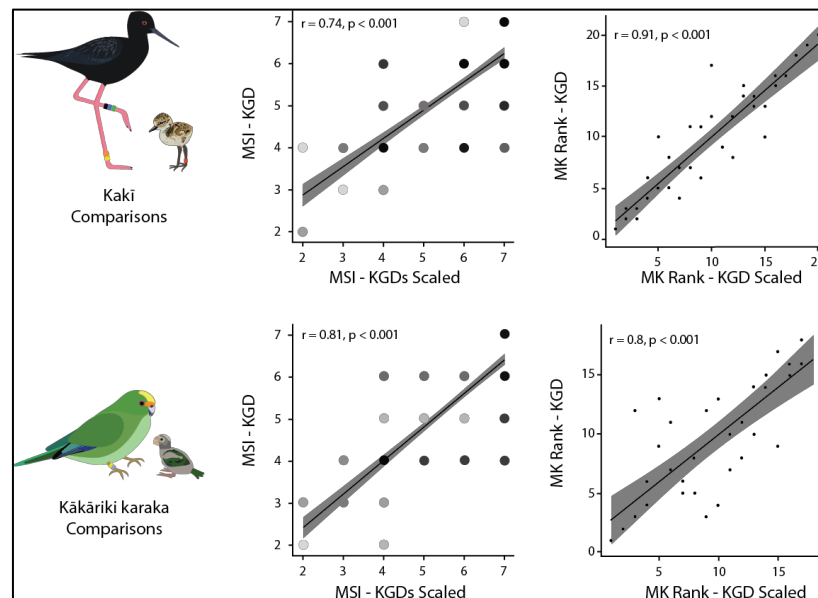


Figure G 3. Scatterplots showing relationships between scaled and unscaled KGD estimates in regards to MSI scores and MK rank in kakī and kākārīki karaka. Darker points denote higher frequencies than lighter points. A trend line (black) and 95% confidence intervals (grey) are shown in each comparison.

## G.5: References:

- Campbell IM, Gambin T, Jhangiani S, Grove ML, Veeraraghavan N, Muzn DM, Shaw CA, Gibbs RA, Boerwinkle E, Yu F, Lupski JR (2016) Multiallelic positions in the human genome: challenges for genetic analyses. *Human Mutations*, 37(3), 231-234.
- Dodds KG, McEwan JC, Brauning R, Anderson RM, van Stijn TC, Kristjánsson T, Clarke SM (2015) Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, **16**, 1047.
- Hammerly SC, Morrow ME, Johnson JA (2013) A comparison of pedigree- and DNA-based measures for identifying inbreeding depression in the critically endangered Attwater's Prairie-chicken. *Molecular Ecology*, **22**, 5313–5328.

- Li M (2010) identity: Jacquard condensed coefficients of identity. R Package v.0.2-1.
- Pew J, Muir PH, Wang J, Frasier TR (2015) Related: an R package for analysing pairwise relatedness from codominant molecular markers. *Molecular Ecology Resources*, **15**, 557-561
- Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, **16**, 33–44.
- Wang J (2007) Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetics Research*, **89**, 135-153.
- Wang J (2011) COANCESTRY: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, **11**, 141-145.

# Appendix H: Chapter 5, SNP Discovery Pipeline and R-code

## H.1: SNP Discovery Pipeline for Kākāriki Karaka

### H.1.2: FastQC

Available at:

[https://github.com/sgalla32/OFK\\_Resequenc\\_DataProcessing/blob/master/OFK\\_FastQC](https://github.com/sgalla32/OFK_Resequenc_DataProcessing/blob/master/OFK_FastQC)

```
#!/bin/sh
```

```
/home/stephanie/FastQC/fastqc /media/stephanie/Data/Resequencing/OFK/I00578-  
L1_S4_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00578-L1_S4_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00579-L1_S5_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00579-L1_S5_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00580-L1_S6_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00580-L1_S6_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00581-L1_S7_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00581-L1_S7_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00582-L1_S8_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00582-L1_S8_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00583-L1_S9_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00583-L1_S9_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00584-L1_S10_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00584-L1_S10_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00585-L1_S11_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00585-L1_S11_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00586-L1_S12_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00586-L1_S12_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00587-L1_S13_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00587-L1_S13_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00588-L1_S14_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00588-L1_S14_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00589-L1_S15_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00589-L1_S15_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00590-L1_S16_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00590-L1_S16_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00591-L1_S17_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00591-L1_S17_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00592-L1_S18_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00592-L1_S18_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00593-L1_S19_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00593-L1_S19_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00594-L1_S20_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00594-L1_S20_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00595-L1_S21_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00595-L1_S21_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00596-L1_S22_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00596-L1_S22_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00597-L1_S23_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00597-L1_S23_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00598-L1_S24_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00598-L1_S24_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00599-L1_S25_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00599-L1_S25_L004_R2_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00600-L1_S26_L004_R1_001.fastq.gz \  
/media/stephanie/Data/Resequencing/OFK/I00600-L1_S26_L004_R2_001.fastq.gz \  

```

```

/media/stephanie/Data/Resequencing/OFK/I00601-L1_S27_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00601-L1_S27_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00602-L1_S28_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00602-L1_S28_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00603-L1_S29_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00603-L1_S29_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00604-L1_S30_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00604-L1_S30_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00605-L1_S31_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00605-L1_S31_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00606-L1_S32_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00606-L1_S32_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00607-L1_S33_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00607-L1_S33_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00608-L1_S34_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00608-L1_S34_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00609-L1_S35_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00609-L1_S35_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00610-L1_S36_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00610-L1_S36_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00611-L1_S37_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00611-L1_S37_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00612-L1_S38_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00612-L1_S38_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00613-L1_S39_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00613-L1_S39_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00614-L1_S40_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00614-L1_S40_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00615-L1_S41_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00615-L1_S41_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00616-L1_S42_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00616-L1_S42_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00617-L1_S43_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00617-L1_S43_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00618-L1_S44_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00618-L1_S44_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00619-L1_S45_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00619-L1_S45_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00620-L1_S46_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00620-L1_S46_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00621-L1_S47_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00621-L1_S47_L004_R2_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00622-L1_S48_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/I00622-L1_S48_L004_R2_001.fastq.gz

```

### H.1.3: TrimGalore

Available at:

[https://github.com/sgalla32/OFK\\_Resequenc\\_DataProcessing/blob/master/OFK\\_TrimGalore](https://github.com/sgalla32/OFK_Resequenc_DataProcessing/blob/master/OFK_TrimGalore)

```
#!/bin/sh
```

```

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00578 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00578-
L1_S4_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00578-
L1_S4_L004_R2_001.fastq.gz

```

```

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00579 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \

```

```

/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00579-
L1_S5_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00579-
L1_S5_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00580 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00580-
L1_S6_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00580-
L1_S6_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00581 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00581-
L1_S7_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00581-
L1_S7_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00582 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00582-
L1_S8_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00582-
L1_S8_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00583 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00583-
L1_S9_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00583-
L1_S9_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00584 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00584-
L1_S10_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00584-
L1_S10_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00585 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00585-
L1_S11_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00585-
L1_S11_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00586 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00586-
L1_S12_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00586-
L1_S12_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00587 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00587-
L1_S13_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00587-
L1_S13_L004_R2_001.fastq.gz

```

```

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00588 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00588-
L1_S14_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00588-
L1_S14_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00589 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00589-
L1_S15_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00589-
L1_S15_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00590 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00590-
L1_S16_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00590-
L1_S16_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00591 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00591-
L1_S17_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00591-
L1_S17_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00592 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00592-
L1_S18_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00592-
L1_S18_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00593 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00593-
L1_S19_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00593-
L1_S19_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00594 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00594-
L1_S20_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00594-
L1_S20_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00595 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00595-
L1_S21_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00595-
L1_S21_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00596 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \

```

```

/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00596-
L1_S22_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00596-
L1_S22_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00597 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00597-
L1_S23_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00597-
L1_S23_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00598 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00598-
L1_S24_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00598-
L1_S24_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00599 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00599-
L1_S25_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00599-
L1_S25_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00600 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00600-
L1_S26_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00600-
L1_S26_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00601 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00601-
L1_S27_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00601-
L1_S27_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00602 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00602-
L1_S28_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00602-
L1_S28_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00603 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00603-
L1_S29_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00603-
L1_S29_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00604 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00604-
L1_S30_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00604-
L1_S30_L004_R2_001.fastq.gz

```

```

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00605 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00605-
L1_S31_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00605-
L1_S31_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00606 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00606-
L1_S32_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00606-
L1_S32_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00607 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00607-
L1_S33_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00607-
L1_S33_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00608 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00608-
L1_S34_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00608-
L1_S34_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00609 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00609-
L1_S35_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00609-
L1_S35_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00610 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00610-
L1_S36_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00610-
L1_S36_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00611 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00611-
L1_S37_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00611-
L1_S37_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00612 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00612-
L1_S38_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00612-
L1_S38_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00613 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \

```



```

/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00613-
L1_S39_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00613-
L1_S39_L004_R2_001.fastq.gz

/home/stephanie/TrimGalore-master/trim_galore --paired --trim-n --nextera --nextseq
20 --quality 30 --length 54 --basename OFK_pe_I00614 --o
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/ \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00614-
L1_S40_L004_R1_001.fastq.gz \
/media/stephanie/Data/Resequencing/OFK/Raw_Data_and_FastQCs/I00614-
L1_S40_L004_R2_001.fastq.gz

```

## H.1.4: Bowtie2

Available at:

[https://github.com/sgalla32/OFK\\_Resequenc\\_DataProcessing/blob/master/Bowtie2\\_Samtools\\_BCFTools](https://github.com/sgalla32/OFK_Resequenc_DataProcessing/blob/master/Bowtie2_Samtools_BCFTools)

```

cd /data/trimmed_fastq

bowtie2-build /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta
/data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00578-
L1_S4_L004_R1_001_val_1.fq.gz -2 I00578-L1_S4_L004_R2_001_val_2.fq.gz -X 600 -I 0 |
samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00578.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00580-
L1_S6_L004_R1_001_val_1.fq.gz -2 I00580-L1_S6_L004_R2_001_val_2.fq.gz -X 600 -I 0 |
samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00580.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00581-
L1_S7_L004_R1_001_val_1.fq.gz -2 I00581-L1_S7_L004_R2_001_val_2.fq.gz -X 600 -I 0 |
samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00581.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00582-
L1_S8_L004_R1_001_val_1.fq.gz -2 I00582-L1_S8_L004_R2_001_val_2.fq.gz -X 600 -I 0 |
samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00582.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00583-
L1_S9_L004_R1_001_val_1.fq.gz -2 I00583-L1_S9_L004_R2_001_val_2.fq.gz -X 600 -I 0 |
samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00583.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00584-
L1_S10_L004_R1_001_val_1.fq.gz -2 I00584-L1_S10_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00584.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00585-
L1_S11_L004_R1_001_val_1.fq.gz -2 I00585-L1_S11_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00585.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00586-
L1_S12_L004_R1_001_val_1.fq.gz -2 I00586-L1_S12_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00586.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00587-
L1_S13_L004_R1_001_val_1.fq.gz -2 I00587-L1_S13_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00587.bam

```

```

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00588-
L1_S14_L004_R1_001_val_1.fq.gz -2 I00588-L1_S14_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00588.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00589-
L1_S15_L004_R1_001_val_1.fq.gz -2 I00589-L1_S15_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00589.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00590-
L1_S16_L004_R1_001_val_1.fq.gz -2 I00590-L1_S16_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00590.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00591-
L1_S17_L004_R1_001_val_1.fq.gz -2 I00591-L1_S17_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00591.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00592-
L1_S18_L004_R1_001_val_1.fq.gz -2 I00592-L1_S18_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00592.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00593-
L1_S19_L004_R1_001_val_1.fq.gz -2 I00593-L1_S19_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00593.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00594-
L1_S20_L004_R1_001_val_1.fq.gz -2 I00594-L1_S20_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00594.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00595-
L1_S21_L004_R1_001_val_1.fq.gz -2 I00595-L1_S21_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00595.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00596-
L1_S22_L004_R1_001_val_1.fq.gz -2 I00596-L1_S22_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00596.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00597-
L1_S23_L004_R1_001_val_1.fq.gz -2 I00597-L1_S23_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00597.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00598-
L1_S24_L004_R1_001_val_1.fq.gz -2 I00598-L1_S24_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00598.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00599-
L1_S25_L004_R1_001_val_1.fq.gz -2 I00599-L1_S25_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00599.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00600-
L1_S26_L004_R1_001_val_1.fq.gz -2 I00600-L1_S26_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00600.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00601-
L1_S27_L004_R1_001_val_1.fq.gz -2 I00601-L1_S27_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00601.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00602-
L1_S28_L004_R1_001_val_1.fq.gz -2 I00602-L1_S28_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00602.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00603-
L1_S29_L004_R1_001_val_1.fq.gz -2 I00603-L1_S29_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00603.bam

```

```

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00604-
L1_S30_L004_R1_001_val_1.fq.gz -2 I00604-L1_S30_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00604.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00605-
L1_S31_L004_R1_001_val_1.fq.gz -2 I00605-L1_S31_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00605.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00606-
L1_S32_L004_R1_001_val_1.fq.gz -2 I00606-L1_S32_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00606.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00607-
L1_S33_L004_R1_001_val_1.fq.gz -2 I00607-L1_S33_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00607.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00608-
L1_S34_L004_R1_001_val_1.fq.gz -2 I00608-L1_S34_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00608.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00609-
L1_S35_L004_R1_001_val_1.fq.gz -2 I00609-L1_S35_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00609.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00610-
L1_S36_L004_R1_001_val_1.fq.gz -2 I00610-L1_S36_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00610.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00611-
L1_S37_L004_R1_001_val_1.fq.gz -2 I00611-L1_S37_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00611.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00612-
L1_S38_L004_R1_001_val_1.fq.gz -2 I00612-L1_S38_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00612.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00613-
L1_S39_L004_R1_001_val_1.fq.gz -2 I00613-L1_S39_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00613.bam

bowtie2 -p 32 -x /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -1 I00614-
L1_S40_L004_R1_001_val_1.fq.gz -2 I00614-L1_S40_L004_R2_001_val_2.fq.gz -X 600 -I 0
| samtools view -@ 4 -bS - > /data/unsorted_bam/OFK_pe_I00614.bam

```

### *H.1.5: Samtools v. 1.9 Sort*

Available at:

[https://github.com/sgalla32/OFK\\_Resequenc\\_DataProcessing/blob/master/Bowtie2\\_Samtools\\_BCFTools](https://github.com/sgalla32/OFK_Resequenc_DataProcessing/blob/master/Bowtie2_Samtools_BCFTools)

```
cd /data/unsorted_bam
```

```
for sample in *
do samtools sort -T /data/sorted_BAM/$sample -o /data/sorted_BAM/$sample $sample
```

```
done
```

## H.1.6: Mpileup

Available at:

[https://github.com/sgalla32/OFK\\_Resequenc\\_DataProcessing/blob/master/Bowtie2\\_Samtools\\_BCFTools](https://github.com/sgalla32/OFK_Resequenc_DataProcessing/blob/master/Bowtie2_Samtools_BCFTools)

```
#!/bin/sh

#####
#split_bamfiles_tasks.pl
#####

perl split_bamfiles_tasks.pl -B -b /scripts/Split_Bamfiles/OFK_sorted_BAMfiles.txt
-g /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -n 12 -o
/data/Split_Bamfiles > list_of_tasks.sh

bash list_of_tasks.sh

#####
#Samtools -mpileup
#####

/data/mpileup_1June2019/List_OFK_unsplit_bams.txt

bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/1/part1.bcf -O bcf -b
/data/Split_Bamfiles/1/OFK.1.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/2/part2.bcf -O bcf -b
/data/Split_Bamfiles/2/OFK.2.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/3/part3.bcf -O bcf -b
/data/Split_Bamfiles/3/OFK.3.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/4/part4.bcf -O bcf -b
/data/Split_Bamfiles/4/OFK.4.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/5/part5.bcf -O bcf -b
/data/Split_Bamfiles/5/OFK.5.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/6/part6.bcf -O bcf -b
/data/Split_Bamfiles/6/OFK.6.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/7/part7.bcf -O bcf -b
/data/Split_Bamfiles/7/OFK.7.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/8/part8.bcf -O bcf -b
/data/Split_Bamfiles/8/OFK.8.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/9/part9.bcf -O bcf -b
/data/Split_Bamfiles/9/OFK.9.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/10/part10.bcf -O bcf -b
/data/Split_Bamfiles/10/OFK.10.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/11/part11.bcf -O bcf -b
/data/Split_Bamfiles/11/OFK.11.list &
bcftools mpileup -f /data/OFK_Masurca_Concat/ofk.masurca.corrl.concat.fasta -a
DP,AD,ADF,ADR,SP -o /data/mpileup_bcfs/12/part12.bcf -O bcf -b
/data/Split_Bamfiles/12/OFK.12.list
```

## H.2.1: BCFTools v. 1.9 View & Prune

Available at:

[https://github.com/sgalla32/OFK\\_Resequenc\\_DataProcessing/blob/master/BCFTools\\_View\\_Concat\\_Prune](https://github.com/sgalla32/OFK_Resequenc_DataProcessing/blob/master/BCFTools_View_Concat_Prune)

```
#!/usr/bin/bash
```

```
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part1.call.bcf -o
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part1.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part2.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part2_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part3.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part3_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part4.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part4_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part5.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part5_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part6.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part6_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part7.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part7_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part8.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part8_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part9.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part9_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part10.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part10_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part11.call.bcf | grep -v \# >>
```

```

/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part11_concatenated.call.vcf
bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part12.call.bcf | grep -v \# >>
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part12_concatenated.call.vcf

```

```

#####
#Cat
#####

```

```

cat
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/part1.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part2_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part3_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part4_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part5_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part6_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part7_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part8_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part9_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part10_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part11_concatenated.call.vcf
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/Part12_concatenated.call.vcf >
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/OFK_Concat_4July2019.vcf

```

```

#####
#Filter&Prune
#####

```

```

#-v snps selects for SNPs only.
# -m2 -M2 selects for biallelic SNPs.
# -q 0.05:minor is an MAF cutoff of 0.05 (could also use the filter MAF option, but
this seems to work well)
# MIN(FMT/DP)>5 is minimum depth of 5.
# MAX(MFT/DP)<200 is a maximum depth of 200.
# %QUAL>20 filters for quality above 20.

```

```

bcftools view
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/OFK_Concat_4July2019.vcf -v snps -m 2 -M 2 -q 0.05:minor -i 'MIN(FMT/DP)>5
& MAX(FMT/DP)<200 & %QUAL>20' -o
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/OFK_Concat_BCFTools_Biallelic_SNPsOnly_MAF0.05_MINDP5_MAXDP200_Q20_4July201
9.vcf

```

```

#prunes for LD R2 of 0.6 with a window of 1000bp
#-e excludes sites that have more than 0.1 missingness.

```

```

bcftools +prune -l 0.6 -w 1000 -e "F_MISSING>0.1"
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par

```

```
ts_1_12/OFK_Concat_BCFTTools_Biallelic_SNPsOnly_MAF0.05_MINDP5_MAXDP200_Q20_4July2019.vcf -Ov -o
/media/stephanie/Data/Resequencing/OFK/Roger_SNPCalling/Part_1_12_BCFs/Original_Par
ts_1_12/OFK_Concat_BCFTTools_Biallelic_SNPsOnly_MAF0.05_MINDP5_MAXDP200_Q20_LD0.6_Mi
ssing0.1_4July2019.vcf
```

## H.2: SNP Filtering using BCFTools v. 1.9 for Kaki

Available at:

[https://github.com/sgalla32/Resequencing\\_Data\\_Processing/blob/master/Filtering\\_Chapter5](https://github.com/sgalla32/Resequencing_Data_Processing/blob/master/Filtering_Chapter5)

```
#Stephanie Galla
#22April2019

#Resequencing Filtering
#-v snps selects for SNPs only.
# -m2 -M2 selects for biallelic SNPs.
# -q 0.05:minor is an MAF cutoff of 0.05 (could also use the filter MAF option, but
this seems to work well)
# AVG(FMT/DP)>10 is an average minimum depth of 10.
# %QUAL>20 filters for quality above 20.

bcftools view Kaki_VariantCalls.bcf -v snps -m 2 -M 2 -q 0.05:minor -i
'AVG(FMT/DP)>10 & %QUAL>20' -o
Kaki_FinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf

#prunes for LD R2 of 0.6 with a window of 1000bp
#-e excludes sites that have more than 0.1 missingness.

bcftools +prune -l 0.6 -w 1000 -e "F_MISSING>=0.1"
Kaki_FinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20.vcf -Ov -o
Kaki_FinalVariantCalls_BCFTTools_Biallelic_SNPsOnly_MAF0.05_AVGDP10_Q20_LD0.6_Missin
g0.1.vcf
```

## H.3: R-Scripts:

### H.3.1: General Script for KGD and Relatedness Scale Transformation

KGD can be downloaded from this repository: <https://github.com/AgResearch/KGD>

```
#If you are converting a vcf from Tassel5 or STACKS, remember to convert the vcf to
an ra.tab file using vcf2ra_ro_ao.py
rm(list = ls())
#Set your working directory for your file locations.
setwd("FILE_LOCATION")
genofile <- "resequencing.full_dataset.vcf.ra.tab"
gform <- "Tassel"
source("GBS-Chip-Gmatrix.R")
Gfull <- calcG()
GHWdgm.05 <- calcG(which(HWdis > -0.05),"HWdgm.05", npc = 4)
str(GHWdgm.05)
#Write out your relatedness estimates in vector form.
writeG(Gfull, "Killdeer_BCFTools", outtype = 3)
#Write out your relatedness estimates in matrix form.
writeG(Gfull, "Killdeer_BCFTools_Matrix", outtype = 2, ,seqID)

#####
#Scale Transformation
#####
```

```

setwd("FILE_LOCATION")

#Removes everything from the R environment. Only use when fresh start needed.
rm(list=ls())

#Reads your csv relatedness matrix into R.
M <- as.matrix(read.csv("OFK_WriteG_Gfull.LD0.6_NoNames.csv", header = FALSE))

#The function diag(x) extracts or replaces the diagonal of a matrix, or construct a
diagonal matrix.
#x can specify a matrix, when it extracts the diagonal.
#x can be missing and nrow is specified, it returns an identity matrix
#x can be scalar (length-one vector) and the only argument it returns a square
identity matrix of size given by the scalar.
#x can be a 'numeric' (complex, numeric, integer, logical, or raw) vector, either
of length at least 2 or there were further arguments. This returns a matrix with
the given diagonal and zero off-diagonal entries.
D <- diag(36) * 1/sqrt(diag(M))

#This creates a scaled matrix with ones as the diagonals.
#Note, % is an infix operator that does math in the background (in this case,
multiplying).
ScaledMatrix <- D %*% M %*% D

#Converts a covariate matrix to a correlation matrix.
ScaledMatrix_Correlation <- cov2cor(ScaledMatrix)

write.table(ScaledMatrix_Correlation, file="OFK_WriteG_Scaled_LD0.6.txt", sep= " ")

write.csv(ScaledMatrix_Correlation, file="OFK_WriteG_Scaled_LD0.6.csv")

```

### ***H.3.2: Boxplots for Parent-Offspring and Sibling Relatedness***

```

#20July2019
#SGalla
#BoxPlots for chapter 5, manuscript 3

#Removes everything from the environment
rm(list=ls())

library("ggplot2",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")

#####
#Kaki
#####

setwd("~/Desktop/Chapter5/Ped_Micro_Reseq_Compare/Kaki/PO_and_Sibs")

#Create a dataset "Kaki_PO_Relatedness" which has your csv, delineated as a tab-
separated file or csv with headers.
Kaki_PO_Relatedness <- read.csv("Kaki_Parent_Offspring_Sibling_Comparison_PO.txt",
sep = "\t", header = TRUE)

#Lets you check to make sure your columns are correctly labeled.
head(Kaki_PO_Relatedness)

#Creates a box plot
ggplot(Kaki_PO_Relatedness, aes(x = Data_Type, y = Relatedness)) + geom_boxplot()

#Create a dataset "Kaki_Sib_Relatedness" which has your csv, delineated as a tab-
separated file or csv with headers.
Kaki_Sib_Relatedness <-
read.csv("Kaki_Parent_Offspring_Sibling_Comparison_Sibs.txt", sep = "\t", header =
TRUE)

```



```

#Lets you check to make sure your columns are correctly labeled.
head(Kaki_Sib_Relatedness)

#Creates a box plot
ggplot(Kaki_Sib_Relatedness, aes(x = Data_Type, y = Relatedness)) + geom_boxplot()

#####
#OFK
#####

#Removes everything from the environment
rm(list=ls())

library("ggplot2",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")

setwd("~/Desktop/Chapter5/Ped_Micro_Reseq_Compare/OFK/PO_and_Sibs")

#Create a dataset "Kaki_PO_Relatedness" which has your csv, delineated as a tab-
separated file or csv with headers.
OFK_PO_Relatedness <- read.csv("OFK_Parent_Offspring_Comparison.csv", sep = ",",
header = TRUE)

#Lets you check to make sure your columns are correctly labeled.
head(OFK_PO_Relatedness)

#Creates a box plot
ggplot(OFK_PO_Relatedness, aes(x = Data_Type, y = Relatedness)) + geom_boxplot()

#Create a dataset "Kaki_Sib_Relatedness" which has your csv, delineated as a tab-
separated file or csv with headers.
OFK_Sib_Relatedness <- read.csv("OFK_Sibling_Comparison.csv", sep = ",", header =
TRUE)

#Lets you check to make sure your columns are correctly labeled.
head(OFK_Sib_Relatedness)

#Creates a box plot
ggplot(OFK_Sib_Relatedness, aes(x = Data_Type, y = Relatedness)) + geom_boxplot()

```

### *H.3.3: Relatedness Correlations*

```

#20July2019
#SGalla
#Pearson's R for Chapter 5

#Clear environment, if needed.
rm(list=ls())

library("ggpubr")

#####
#Kaki
#####

#Set your working directory
setwd("~/Desktop/Chapter5/Resequencing/Kaki/Variant_Filtering_Trials/BaseFiltering_
LD0.6/Kaki/SNPS_KGD_v_TrioML")

#To input your resequencing KGD vs. COANCESTRY TrioML run:
# If .txt or csv file, use this
Kaki_KGD_Coancestry_Reseq <-
read.csv("Kaki_Base_Filtering_LD0.6_COANCESTRY_v_KGD.csv", sep = ",", header =
TRUE)
head(Kaki_KGD_Coancestry_Reseq, 6)

```

```

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll
use the ggpubr R package.
ggscatter(Kaki_KGD_Coancestry_Reseq, x = "SNPs_COANCESTRY_R", y =
"KGD_Scaled_LD0.6",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "TrioML", ylab = "KGD")

#Is the covariation linear? Yes, from the plot above, the relationship is linear.
#In the situation where the scatter plots show curved patterns, we are dealing with
nonlinear association between the two variables.
#Are the data from each of the 2 variables (x, y) follow a normal distribution?

sink('Kaki_shapiro-wilk_normality_SNPs_KGD_TrioML.txt')
# Shapiro-Wilk normality test for TrioML
shapiro.test(Kaki_KGD_Coancestry_Reseq$SNPs_COANCESTRY_R)
# Shapiro-Wilk normality test for KGD
shapiro.test(Kaki_KGD_Coancestry_Reseq$KGD_Scaled_LD0.6)

sink()

#Visual inspection of the data normality using Q-Q plots (quantile-quantile
plots).
#Q-Q plot draws the correlation between a given sample and the normal distribution.

# TrioML
ggqqplot(Kaki_KGD_Coancestry_Reseq$SNPs_COANCESTRY_R, ylab = "TrioML")
# KGD
ggqqplot(Kaki_KGD_Coancestry_Reseq$KGD_Scaled, ylab = "KGD")

#Correlation tests:
sink('Kaki_Correlation_SNP_TrioMLvsKGD.txt')

Kaki_SNP_TrioMLvsKGD_Pearsons <-
cor.test(Kaki_KGD_Coancestry_Reseq$SNPs_COANCESTRY_R,
          Kaki_KGD_Coancestry_Reseq$KGD_Scaled_LD0.
6,
          method = "pearson")

Kaki_SNP_TrioMLvsKGD_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Kaki_SNP_TrioMLvsKGD_Kendall <-
cor.test(Kaki_KGD_Coancestry_Reseq$SNPs_COANCESTRY_R,
          Kaki_KGD_Coancestry_Reseq$KGD_Scaled_LD0.6
,
          method = "kendall")

Kaki_SNP_TrioMLvsKGD_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Kaki_SNP_TrioMLvsKGD_Spearman <-
cor.test(Kaki_KGD_Coancestry_Reseq$SNPs_COANCESTRY_R,
          Kaki_KGD_Coancestry_Reseq$KGD_Scaled_LD0.
6,
          method = "spearman")

Kaki_SNP_TrioMLvsKGD_Spearman

sink()

#####

#Comparing Pedigree, Genetic, and Genomic-based R via correlations

#Set your working directory
setwd("~/Desktop/Chapter5/Ped_Micro_Reseq_Compare/Kaki/Correlations_All/")

```

```

#To input your resequencing KGD vs. COANCESTRY TrioML run:
# If .txt or csv file, use this
Kaki_Comparison_Ped_Mic_SNPs <- read.csv("Comparison_Ped_Micro_Reseq_NoInds.csv",
sep = ",", header = TRUE)
head(Kaki_Comparison_Ped_Mic_SNPs, 6)

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll
use the ggpubr R package.
ggscatter(Kaki_Comparison_Ped_Mic_SNPs, x = "Pedigree", y = "Genetic.TrioML",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Pedigree", ylab = "TrioML")

ggscatter(Kaki_Comparison_Ped_Mic_SNPs, x = "Pedigree", y = "KGD_Scaled_LD0.6",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Pedigree", ylab = "KGD_Scaled_LD0.6")

ggscatter(Kaki_Comparison_Ped_Mic_SNPs, x = "Genetic.TrioML", y =
"KGD_Scaled_LD0.6",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "TrioML", ylab = "KGD_Scaled_LD0.6")

#Is the covariation linear? Yes, form the plot above, the relationship is linear.
#In the situation where the scatter plots show curved patterns, we are dealing with
nonlinear association between the two variables.
#Are the data from each of the 2 variables (x, y) follow a normal distribution?

sink('Kaki_shapiro-wilk_normality_All_Relatedness.txt')
# Shapiro-Wilk normality test for pedigree
shapiro.test(Kaki_Comparison_Ped_Mic_SNPs$Pedigree)
# Shapiro-Wilk normality test for micros
shapiro.test(Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML)
# Shapiro-Wilk normality test for SNPs
shapiro.test(Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6)

sink()

#Visual inspection of the data normality using Q-Q plots (quantile-quantile
plots).
#Q-Q plot draws the correlation between a given sample and the normal distribution.

# Pedigree
ggqqplot(Kaki_Comparison_Ped_Mic_SNPs$Pedigree, ylab = "Pedigree")
# Micro
ggqqplot(Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML, ylab = "Micro_TrioML")
# SNPs
ggqqplot(Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6, ylab = "SNP_KGD")

#Pearson's Correlation test between F and Reference variables:
sink('Kaki_Correlation_All_Relatedness.txt')

Kaki_Ped_Micro_Pearsons <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Pedigree,
                                   Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML,
                                   method = "pearson")

Kaki_Ped_Micro_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Kaki_Ped_Micro_Kendall <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Pedigree,
                                   Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML,
                                   method = "kendall")

Kaki_Ped_Micro_Kendall

#Spearman Rank Correlation test (for non-parametric data)

```

```

Kaki_Ped_Micro_Spearman <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Pedigree,
                                     Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML,
                                     method = "spearman")

Kaki_Ped_Micro_Spearman

Kaki_Ped_SNPs_Pearsons <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Pedigree,
                                    Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6,
                                    method = "pearson")

Kaki_Ped_SNPs_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Kaki_Ped_SNPs_Kendall <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Pedigree,
                                   Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6,
                                   method = "kendall")

Kaki_Ped_SNPs_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Kaki_Ped_SNPs_Spearman <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Pedigree,
                                    Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6,
                                    method = "spearman")

Kaki_Ped_SNPs_Spearman

Kaki_Micro_SNPs_Pearsons <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML,
                                      Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6,

                                      method = "pearson")

Kaki_Micro_SNPs_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Kaki_Micro_SNPs_Kendall <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML,
                                    Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6,
                                    method = "kendall")

Kaki_Micro_SNPs_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Kaki_Micro_SNPs_Spearman <- cor.test(Kaki_Comparison_Ped_Mic_SNPs$Genetic.TrioML,
                                      Kaki_Comparison_Ped_Mic_SNPs$KGD_Scaled_LD0.6,

                                      method = "spearman")

Kaki_Micro_SNPs_Spearman

sink()

#####
#OFK
#####

library("ggpubr")

setwd("~/Desktop/Chapter5/Ped_Micro_Reseq_Compare/OFK/Correlations_All")

#To input your resequencing KGD vs. COANCESTRY TrioML run:
# If .txt or csv file, use this
OFK_KGD_Coancestry_Reseq <- read.csv("All_OFK_Comparison_Ped_Micro_Reseq.csv", sep
= ",", header = TRUE)
head(OFK_KGD_Coancestry_Reseq, 6)

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll
use the ggpubr R package.

```

```

ggscatter(OFK_KGD_Coancestry_Reseq, x = "TrioML", y = "SNPs",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "TrioML", ylab = "KGD")

#Is the covariation linear? Yes, form the plot above, the relationship is linear.
#In the situation where the scatter plots show curved patterns, we are dealing with
nonlinear association between the two variables.
#Are the data from each of the 2 variables (x, y) follow a normal distribution?

sink('OFK_shapiro-wilk_normality_SNPs_KGD_TrioML.txt')
# Shapiro-Wilk normality test for TrioML
shapiro.test(OFK_KGD_Coancestry_Reseq$TrioML)
# Shapiro-Wilk normality test for KGD
shapiro.test(OFK_KGD_Coancestry_Reseq$SNPs)
sink()

#Visual inspection of the data normality using Q-Q plots (quantile-quantile
plots).
#Q-Q plot draws the correlation between a given sample and the normal distribution.

# TrioML
ggqqplot(OFK_KGD_Coancestry_Reseq$TrioML, ylab = "TrioML")
# KGD
ggqqplot(OFK_KGD_Coancestry_Reseq$SNPs, ylab = "KGD")

#Correlation tests
sink('OFK_Correlation_SNP_TrioMLvsKGD.txt')

OFK_SNP_TrioMLvsKGD_Pearsons <- cor.test(OFK_KGD_Coancestry_Reseq$TrioML,
                                       OFK_KGD_Coancestry_Reseq$SNPs,
                                       method = "pearson")

OFK_SNP_TrioMLvsKGD_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
OFK_SNP_TrioMLvsKGD_Kendall <- cor.test(OFK_KGD_Coancestry_Reseq$TrioML,
                                       OFK_KGD_Coancestry_Reseq$SNPs,
                                       method = "kendall")

OFK_SNP_TrioMLvsKGD_Kendall

#Spearman Rank Correlation test (for non-parametric data)
OFK_SNP_TrioMLvsKGD_Spearman <- cor.test(OFK_KGD_Coancestry_Reseq$TrioML,
                                       OFK_KGD_Coancestry_Reseq$SNPs,
                                       method = "spearman")

OFK_SNP_TrioMLvsKGD_Spearman

sink()

#####

OFK_Comparison <- read.csv("All_OFK_Comparison_Ped_Micro_Reseq.csv", sep = ",",
header = TRUE)
head(OFK_Comparison, 6)

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll
use the ggpubr R package.
ggscatter(OFK_Comparison, x = "Pedigree", y = "Micro",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Pedigree", ylab = "Micro")

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll
use the ggpubr R package.
ggscatter(OFK_Comparison, x = "Pedigree", y = "SNPs",
          add = "reg.line", conf.int = TRUE,

```

```

cor.coef = TRUE, cor.method = "pearson",
xlab = "Pedigree", ylab = "SNPs")

#To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll
use the ggpubr R package.
ggscatter(OFK_Comparison, x = "Micro", y = "SNPs",
add = "reg.line", conf.int = TRUE,
cor.coef = TRUE, cor.method = "pearson",
xlab = "Micro", ylab = "SNPs")

#Is the covariation linear? Yes, from the plot above, the relationship is linear.
#In the situation where the scatter plots show curved patterns, we are dealing with
nonlinear association between the two variables.
#Are the data from each of the 2 variables (x, y) follow a normal distribution?

sink('OFK_shapiro-wilk_normality.txt')
# Shapiro-Wilk normality
shapiro.test(OFK_Comparison$Pedigree)
shapiro.test(OFK_Comparison$Micro)
shapiro.test(OFK_Comparison$SNPs)
sink()

#Visual inspection of the data normality using Q-Q plots (quantile-quantile
plots).
#Q-Q plot draws the correlation between a given sample and the normal distribution.

ggqqplot(OFK_Comparison$Pedigree, ylab = "Pedigree")

ggqqplot(OFK_Comparison$Micro, ylab = "Micro")

ggqqplot(OFK_Comparison$Micro, ylab = "SNP")

#Correlation tests:

sink('Correlation_All_Relatedness.txt')

Ped_Micro_Pearsons <- cor.test(OFK_Comparison$Pedigree,
OFK_Comparison$Micro,
method = "pearson")

Ped_Micro_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Ped_Micro_Kendall <- cor.test(OFK_Comparison$Pedigree,
OFK_Comparison$Micro,
method = "kendall")

Ped_Micro_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Ped_Micro_Spearman <- cor.test(OFK_Comparison$Pedigree,
OFK_Comparison$Micro,
method = "spearman")

Ped_Micro_Spearman

Ped_SNPs_Pearsons <- cor.test(OFK_Comparison$Pedigree,
OFK_Comparison$SNPs,
method = "pearson")

Ped_SNPs_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Ped_SNPs_Kendall <- cor.test(OFK_Comparison$Pedigree,
OFK_Comparison$SNPs,
method = "kendall")

```

```

Ped_SNPs_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Ped_SNPs_Spearman <- cor.test(OFK_Comparison$Pedigree,
                              OFK_Comparison$SNPs,
                              method = "spearman")

Ped_SNPs_Spearman

Micro_SNPs_Pearsons <- cor.test(OFK_Comparison$Micro,
                                OFK_Comparison$SNPs,
                                method = "pearson")

Micro_SNPs_Pearsons

#Kendall Rank Correlation test (for non-parametric data)
Micro_SNPs_Kendall <- cor.test(OFK_Comparison$Micro,
                                OFK_Comparison$SNPs,
                                method = "kendall")

Micro_SNPs_Kendall

#Spearman Rank Correlation test (for non-parametric data)
Micro_SNPs_Spearman <- cor.test(OFK_Comparison$Micro,
                                OFK_Comparison$SNPs,
                                method = "spearman")

Micro_SNPs_Spearman

sink()

```

### ***H.3.4: MSI Correlations***

```

#Stephanie Galla
#20July2019
#Comparison of MSI Scores in Chapter 5.

#Removes everything from the environment
rm(list=ls())

#Check packages for dplyr, pastecs, ggplot2, datasets, graphics, stats, ggpubr
library("dplyr",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")
library("pastecs",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")
library("ggplot2",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")
library("ggpubr",
lib.loc="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")

#####
#Kaki
#####

setwd("~/Desktop/Chapter5/MSI/Kaki/Analyses")

#Create a dataset "MyData" which has your text file, delineated as a tab-separated
file with headers.
Kaki_MSI_Data <- read.csv("Kaki_MSI_2Column.csv", sep = ",", header = TRUE)
Kaki_MSI_Data_Normality <- read.csv("Kaki_MSI_All.csv", sep = ",", header = TRUE)

# Show a random sample to check your data.
set.seed(1234)

```

```

dplyr::sample_n(MSI_Data, 10)

# Lets you check to make sure your columns are correctly labeled.
head(Kaki_MSI_Data)

# Show the levels (or variables that you have)
levels(Kaki_MSI_Data$Approach)

# If the levels are not automatically in the correct order, re-order them as follow:
Kaki_MSI_Data$Approach <- ordered(Kaki_MSI_Data$Approach,
                                  levels = c("Pedigree", "Microsatellite", "SNP"))

# Get min, 1st quartile, median, mean, 3rd quartile, max.
summary(Kaki_MSI_Data_Normality)

# Make sure the package pastecs is clicked. The following line will give you all
# sorts of descriptive stats, and will output it to a new dataset.
stat.desc(Kaki_MSI_Data_Normality, basic = TRUE, desc = TRUE, norm = TRUE, p=0.95)
-> Kaki_DescriptiveStatsMSIData

# Output the results to a txt file
write.table(Kaki_DescriptiveStatsMSIData, "./Kaki_DescriptiveStatsMSIData.txt",
            sep="\t")

sink('Kaki_shapiro-wilk_normality_MSI.txt')
# Shapiro-Wilk normality test for MSI
shapiro.test(Kaki_MSI_Data_Normality$Pedigree)
shapiro.test(Kaki_MSI_Data_Normality$Micro)
shapiro.test(Kaki_MSI_Data_Normality$SNPs)

sink()

ggscatter(Kaki_MSI_Data_Normality, x = "Pedigree", y = "Micro",
          add = "reg.line", conf.int = TRUE, alpha = 0.1,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Pedigree", ylab = "Microsatellite")

ggscatter(Kaki_MSI_Data_Normality, x = "Pedigree", y = "SNPs",
          add = "reg.line", conf.int = TRUE, alpha = 0.1,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Pedigree", ylab = "SNP")

ggscatter(Kaki_MSI_Data_Normality, x = "Micro", y = "SNPs",
          add = "reg.line", conf.int = TRUE, alpha = 0.1,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Microsatellite", ylab = "SNP")

# Creates a box plot
ggplot(Kaki_MSI_Data, aes(x = Approach, y = MSI)) + geom_boxplot()

# Mean plots
# ++++++
# Plot weight by group
# Add error bars: mean_se
# (other values include: mean_sd, mean_ci, median_iqr, ...)
library("ggpubr")
ggline(Kaki_MSI_Data, x = "Approach", y = "MSI",
       add = c("mean_se", "jitter"),
       order = c("Pedigree", "Microsatellite", "SNP"),
       ylab = "MSI", xlab = "Approach")

sink("Kaki_Kruskal_Wallis_Wilcox_MSI.txt")

```



```

#Kruskal Wallis for non-parametric test between R and Reference variables:
kruskal.test(MSI ~ Approach, data = MSI_Data)

#Pairwise Wilcox Test for non-parametric analysis
pairwise.wilcox.test(Kaki_MSI_Data$MSI, Kaki_MSI_Data$Approach,
                     p.adjust.method = "bonferroni")
sink()

#####
#Pearson-Kaki
#####

Comparison_MSI_Kaki <- read.csv("Kaki_MSI_All.csv", sep = ",", header = TRUE)
head(Comparison_MSI_Kaki, 6)

sink('Correlation_29April2019_MSI.txt')

MSI_Pearsons_Pedigree_Micro <- cor.test(Comparison_MSI_Kaki$Pedigree,
                                       Comparison_MSI_Kaki$Micro,
                                       method = "pearson")

MSI_Pearsons_Pedigree_Micro

MSI_Pearsons_Pedigree_SNP <- cor.test(Comparison_MSI_Kaki$Pedigree,
                                       Comparison_MSI_Kaki$SNPs,
                                       method = "pearson")

MSI_Pearsons_Pedigree_SNP

MSI_Pearsons_Micro_SNP <- cor.test(Comparison_MSI_Kaki$Micro,
                                   Comparison_MSI_Kaki$SNPs,
                                   method = "pearson")

MSI_Pearsons_Micro_SNP

sink()

#####
#OFK
#####

setwd("~/Desktop/Chapter5/MSI/OFK/Analyses")

OFK_MSI_Data <- read.csv("OFK_MSI_2Column.csv", sep = ",", header = TRUE)
OFK_MSI_Data_Normality <- read.csv("OFK_MSI_All.csv", sep = ",", header = TRUE)

# Show a random sample to check your data.
set.seed(1234)
dplyr::sample_n(OFK_MSI_Data, 10)

#Lets you check to make sure your columns are correctly labeled.
head(OFK_MSI_Data)

# Show the levels (or variables that you have)
levels(OFK_MSI_Data$Approach)

#If the levels are not automatically in the correct order, re-order them as follow:
OFK_MSI_Data$Approach <- ordered(OFK_MSI_Data$Approach,
                                levels = c("Pedigree", "Micro", "SNPs"))

```

```

#Get min, 1st quartile, median, mean, 3rd quartile, max.
summary(OFK_MSI_Data_Normality)

#Make sure the package pastecs is clicked. The following line will give you all
sorts of descriptive stats, and will output it to a new dataset.
stat.desc(OFK_MSI_Data_Normality, basic = TRUE, desc = TRUE, norm = TRUE, p=0.95) -
> OFK_DescriptiveStatsMSIData

#Output the results to a txt file
write.table(OFK_DescriptiveStatsMSIData, "./OFK_DescriptiveStatsMSIData.txt",
sep="\t")

sink('OFKshapiro-wilk_normality_MSI.txt')
# Shapiro-Wilk normality test for MSI
shapiro.test(OFK_MSI_Data_Normality$Pedigree)
shapiro.test(OFK_MSI_Data_Normality$Micro)
shapiro.test(OFK_MSI_Data_Normality$SNPs)

sink()

ggscatter(OFK_MSI_Data_Normality, x = "Pedigree", y = "Micro",
add = "reg.line", conf.int = TRUE, alpha = 0.1,
cor.coef = TRUE, cor.method = "pearson",
xlab = "Pedigree", ylab = "Microsatellite")

ggscatter(OFK_MSI_Data_Normality, x = "Pedigree", y = "SNPs",
add = "reg.line", conf.int = TRUE, alpha = 0.1,
cor.coef = TRUE, cor.method = "pearson",
xlab = "Pedigree", ylab = "SNP")

ggscatter(OFK_MSI_Data_Normality, x = "Micro", y = "SNPs",
add = "reg.line", conf.int = TRUE, alpha = 0.1,
cor.coef = TRUE, cor.method = "pearson",
xlab = "Microsatellite", ylab = "SNP")

#Creates a box plot
ggplot(OFK_MSI_Data, aes(x = Approach, y = MSI)) + geom_boxplot()

# Mean plots
# ++++++
# Plot weight by group
# Add error bars: mean_se
# (other values include: mean_sd, mean_ci, median_iqr, ....)
library("ggpubr")
ggline(OFK_MSI_Data, x = "Approach", y = "MSI",
add = c("mean_se", "jitter"),
order = c("Pedigree", "Micro", "SNPs"),
ylab = "MSI", xlab = "Approach")

sink("OFK_Kruskal_Wallis_Wilcox_MSI.txt")

#Kruskal Wallis for non-parametric test between R and Reference variables:
kruskal.test(MSI ~ Approach, data = OFK_MSI_Data)

#Pairwise Wilcox Test for non-parametric analysis
pairwise.wilcox.test(OFK_MSI_Data$MSI, OFK_MSI_Data$Approach,
p.adjust.method = "bonferroni")
sink()

#####
#Pearson OFK

```

```
#####

#Set your working directory.
setwd("~/Desktop/Chapter5/MSI/OFK/Analyses")

Comparison_MSI_OFK <- read.csv("OFK_MSI_All.csv", sep = ",", header = TRUE)
head(Comparison_MSI_OFK, 6)

sink('Correlation_OFK_7July2019_MSI.txt')

MSI_Pearsons_Pedigree_Micro <- cor.test(Comparison_MSI_OFK$Pedigree,
                                         Comparison_MSI_OFK$Micro,
                                         method = "pearson")

MSI_Pearsons_Pedigree_Micro

MSI_Pearsons_Pedigree_SNP <- cor.test(Comparison_MSI_OFK$Pedigree,
                                       Comparison_MSI_OFK$SNPs,
                                       method = "pearson")

MSI_Pearsons_Pedigree_SNP

MSI_Pearsons_Micro_SNP <- cor.test(Comparison_MSI_OFK$Micro,
                                   Comparison_MSI_OFK$SNPs,
                                   method = "pearson")

MSI_Pearsons_Micro_SNP

sink()
```